Complex network controllability and applications to biomolecular networks

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Division of Biomedical Engineering
University of Saskatchewan
Saskatoon

By
Lin Wu

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Division of Biomedical Engineering
Room 2B60 Engineering Building
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan Canada S7N 5A9

OR

Dean of College of Graduate and Postdoctoral Studies
Room 116 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan Canada S7N 5C9

Abstract

Within living cells, biomolecules rarely function as isolated elements, but rather interact with each other to perform various biological functions. Biomolecules and their interactions can be represented as biomolecular networks, in which nodes represent different biomolecules and edges represent interactions among biomolecules. The studies of biomolecular networks are critical for understanding the roles of biomolecules within cells and the mechanisms of cellular behaviors. Due to the interactions among biomolecules, the perturbation of some biomolecules may affect others which may eventually change the states of biomolecular networks and corresponding cellular behaviors. Therefore, it is essential and helpful to study biomolecular networks from the viewpoint of control theory.

This thesis investigates the controllability of biomolecular networks based on modern control theory, especially the structural controllability of complex networks. Controllability, which is an important concept in modern control theory, measures the ability of moving a network around in its state space via proper input control signals. To control a complex network, the first step is to identify steering nodes that guarantee the controllability of the network, where steering nodes are nodes directly actuated by input control signals. Although various algorithms have been proposed to identify steering nodes for general complex networks, the applications of the algorithms to biomolecular networks are limited and still have room to be further improved. This thesis focuses on identifying steering nodes, which are biomolecules, for different control scenarios in biomolecular networks.

Three different control scenarios are considered in this thesis. First, to control a network, it's meaningful to determine the least number of nodes which should be actuated by input control signals. To deal with this problem, an algorithm to identify the minimum steering node sets (MSSs) required to have complex networks completely controllable is presented and its applications to biomolecular networks are given. Second, sometimes the minimum number of steering nodes for complete controllability of a network is too large in practical applications. Actually in practice, the complete controllability of all nodes is not necessary. Therefore, an algorithm is developed to identify steering node sets for output controllability, which measures the controllability of a portion of nodes. The algorithm provides a novel method for drug target identification in biomolecular networks: the states of disease related biomolecules can be controlled by actuating the identified steering nodes, which are potential drug targets. Third, to improve feasibility of identified steering nodes in real applications to biomolecular networks, the ability of steering nodes to bind drugs should be considered. An algorithm is proposed to identify steering node sets with drug binding preference. It is expected that steering node sets identified with drug binding preference have more chances to bind to existing drugs, compared to other feasible steering node sets, which facilitates the subsequent procedures of realizing the control of biomolecular networks via drugs.

In addition, in this thesis a software system called CytoCtrlAnalyser is implemented, which contains nine recent algorithms for users to conveniently investigate controllability of biomolecular networks within the Cy-

toscape environment. The algorithms integrated in CytoCtrlAnalyser can be divided into two aspects based on their functions: identifying steering nodes for different control objectives and qualifying or quantifying importance of individual nodes to the controllability of networks.

Keywords: Network controllability, biomolecular networks, steering nodes, realistic control scenarios

Acknowledgements

I owe my deepest gratitude to my supervisor Prof. Fang-Xiang Wu, who introduced me to the world of scientific research. Throughout my study, he has provided me constant encouragement and insightful advices. This thesis would not have been completed without his support and help.

I would also like to express my gratitude to other members of my advisory committee Prof. Tony Kusalik, Prof. Mark Keil and Prof. Jian Yang for their invaluable suggestions and comments during my PhD program.

I am also indebted to my group members Bolin Chen, Lizhi Liu, Yan Yan, Yichao Shen, Amin Mohammadbagheri, Ping Luo and Lingkai Tang for their help in both my life and research.

I would like to thank all my families for their continuous support and love.

Finally, I gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC), University of Saskatchewan (UofS), and the China Scholarship Council (CSC) for the financial supports.

This thesis is dedicated to my father,

Prof. Kun Wu,

who taught me how to be a man;

My mother,

Gongqiu Chen,

who always loves me and supports me;

My wife,

Rui Zhang,

who always gives me strength;

And my daughter,

Emma Jialu Wu,

who brings joy and happiness to my family.

Contents

List of Tables

List of Figures

List of Abbreviations

AAnetwork	acid metabolic network
APL	acute promyelocytic leukemia
CAC	colitis-associated colon cancer
EMT	epithelial to mesenchymal transition
GDCOS	generic dimension of the controllable output subspace
GDCS	generic dimension of the controllable subspace
GPCR	G protein-coupled receptor
GRN	gene regulatory network
JAK3	Janus kinase 3
KEGG	Kyoto Encyclopedia of Genes and Genomes
KM	Kuhn-Munkres
MCMF	minimum cost maximum flow
MDS	minimum driver node set
MRI	magnetic resonance imaging
MSS	minimum steering node set
ODE	ordinary differential equation
PDS	power dominating set
PPI	protein-protein interaction
ROI	region of interest
RTK	receptor tyrosine kinase
SBD	switchboard dynamics
SCC	strong connective component

1 Introduction

1.1 Background

There are various biomolecules in a human body, which include estimated 19,000-20,000 human protein-coding genes [1], ∼1,000 metabolites and an undetermined number of proteins and RNA molecules [2]. Biomolecules perform their functions by interacting with each other and form biomolecular networks, in which nodes represent biomolecules and edges represent the interactions between them. With the exceptional development in high throughput technology, large-scale biological data are available, which provides the foundation for inferring relationships between biomolecules and benefits the reconstructions of different types of large-scale biomolecular networks, such as gene regulatory networks [3], signal transduction networks [4] and protein-protein interaction (PPI) networks [5]. Since diverse cellular functions are typically carried out by the complicated interactions among biomolecules, methods which investigate biomolecular networks systematically are critical for understanding the mechanisms of various biomolecular systems.

Network science, which focuses on analyzing complex relationships between objects mainly based on the topology of networks, has provided powerful methods for understanding underlying relationships among different biomolecules and has prompted valuable progress in drug target identification [6, 7], human disease gene prediction [8], protein complex identification [9], etc. However, studies focusing on topology of biomolecular networks mainly investigate static relationships between biomolecules whereas biomolecular networks represent dynamic systems, in which individual nodes are dynamic and affect each other through interactions. For example, nodes in a gene regulatory network represent genes and each node has a corresponding state variable which represents the expression level of the gene. Then the edges between genes indicate whether the expression of one gene would affect the expression of another gene. For a gene regulatory network, different expression levels of genes indicate different states of the network, which can not be described by network topological structure. Mathematical models have been developed to describe the dynamics of biomolecular systems based on biological observations [10]. A variety of celluar processes have been described successfully [11–13] and lots of promising applications have been made, such as identification of potential drug targets [14–16]. Nevertheless, modeling and analysing the dynamics of large scale biomolecular networks is still challenging due to the difficulties of estimating model parameters and the lack of mathematic tools to analyse the large scale networks.

It has been discovered that many large scale networks of real systems share some non-trivial topological features, such as a heavy tail in the degree distribution, high clustering coefficient, community structure, and hierarchical structure. The networks with these non-trivial topological features are referred to as complex networks and two famous and most studied types of complex networks are scale-free networks [17] and small-world networks [18]. To understand the dynamics of complex networks, a recent pioneering work has investigated the controllability of complex networks [19]. The study [19] used the linear structural dynamic model [20] to describe the dynamics of complex networks. To construct linear structural dynamic models, it is not necessary to estimate the parameters in the models and the models can be determined based on the interaction relationship among nodes. In addition, the mathematic descriptions of the controllability of structural linear models can be formulated as graph-theoretic descriptions, such that the controllability of large networks can be studied by using graph-theoretic algorithms while it is usually formidable for mathematic methods. Taking advantage of linear structural models, the study [19] used the maximum matching algorithm [21] to determine the minimum number of independent input control signals required for having a complex network completely controllable and a minimum driver node set (MDS), in which each node corresponds to an input control signal.

Since many biomolecular networks display many features of complex networks, this thesis investigates the controllability of biomolecular networks based on the linear structural model by extending previous preeminent works or theorems [19, 20, 22–24]. When perturbing some biomolecules in a biomolecular network, the states of other biomolecules may also be affected due to the interactions, which may change the state of the whole biomolecular network. Thus this thesis focuses on identifying the minimum set of steering nodes in biomolecular networks, such that by proper input control signals, the biomolecular networks are controllable and the states of biomolecular networks can be steered to desired ones.

1.2 Motivation and objectives

Our final goal of understanding biomolecular networks is to change their states to what we desire. By changing the states of biomolecular networks, we would be able to modify cellular behaviors or phenotypes. In order to control a network, the main problem is to identify the steering nodes which should be actuated by input control signals. Controllability is a concept in modern control theory, which measures the ability of changing the states of networks by actuating the steering nodes via input control signals. If a network is completely controllable, it can be steered from any initial state to any final state in finite time with appropriate input control signals. Recent progresses have been made to understand the controllability of general complex networks [19, 25, 26]. This thesis focuses on developing methods to identify the minimum set of steering nodes in biomolecular networks such that the biomolecular networks are controllable according to network controllability theorems. To control biomolecular networks, many realistic facts of biomolecular networks should be taken into consideration when identifying the minimum set of steering nodes, such that

the "controllability in principle" can be realized as the "controllability in practice" [27].

Firstly, previous studies mainly focus on complete controllability of the networks and most of the attention have been paid to the minimum driver node set (MDS) proposed by Liu *et. al.* [19]. However, applying independent input control signals to nodes in an MDS respectively is a necessary but not sufficient condition for complete controllability of a network. Therefore, methods are required to identify the minimum steering node sets (MSSs) of biomolecular networks, such that applying independent input control signals to the nodes in an MSS respectively is a sufficient and necessary condition for complete controllability of a network.

Secondly, two practical constraints are considered in the thesis. On the one hand, when controlling biomolecular networks, it is difficult and unnecessary to completely control the whole biomolecular network in many applications. Therefore, algorithms are needed to identify the steering nodes such that the subsets of biomolecules in biomolecular networks can be controlled. Since the biomolecules which are intended to be controlled can be considered as outputs of the biomolecular networks, the problem can be formulated as an output controllability problem. Compared to the MSSs for complete controllability, fewer steering nodes are required for output controllability, which is more practical and efficient in controlling biomolecular networks. On the other hand, since steering node sets for the same control objective are not unique and previous methods of identifying steering node sets only return one possible solution, it could be more feasible if some practical information of steering nodes is considered. To actuate steering nodes in biomolecular networks, input control signals are usually chosen from chemical molecules such as drugs. Therefore, steering node sets which have more chemical-binding opportunities with drugs are more appropriate than randomly chosen steering node sets when actually realizing the control of biomolecular networks by using drugs. To identify steering node sets with drug binding preference, firstly schemes are required for quantifying the drug binding preferences of biomolecules. Then algorithms to identify steering node sets based on preference values are needed. The identification of steering nodes with drug binding preference provides new methods for drug target identification and drug repositioning: steering nodes for the controllability of a disease related network could be potential drug targets while the disease could be a new indication for drugs that can bind to the identified drug targets.

Finally, controlling complex networks has attracted lots of investigation in recent years, so a software system which integrates algorithms of network controllability is in demand. Cytoscape [28] is a software environment for visualizing and analyzing complex networks and a lot of apps, which are actually plug-ins, are available in the Cytoscape App Store for a variety of problems. Therefore, developing the software for network controllability as an app in the Cytoscape environment would be easy to access and use, which would benefit researchers for investigating controllability of biomolecular networks as well as other complex networks.

Based on these motivations, the objectives of this study are described as follows (see Fig. 1.1):

1. Reviewing currently developed methods for controllability of complex networks and their applications to biological networks including biomolecular networks.

2. Developing algorithms for identifying minimum steering node sets for complete controllability of biomolecular networks.

3. Developing algorithms for identifying steering node sets for output controllability of complex networks and applying them to drug target identification in biomolecular networks.

4. Developing algorithms for identifying steering node sets with drug binding preference for controlling biomolecular networks

5. Developing a software system that integrates various network controllability algorithms.

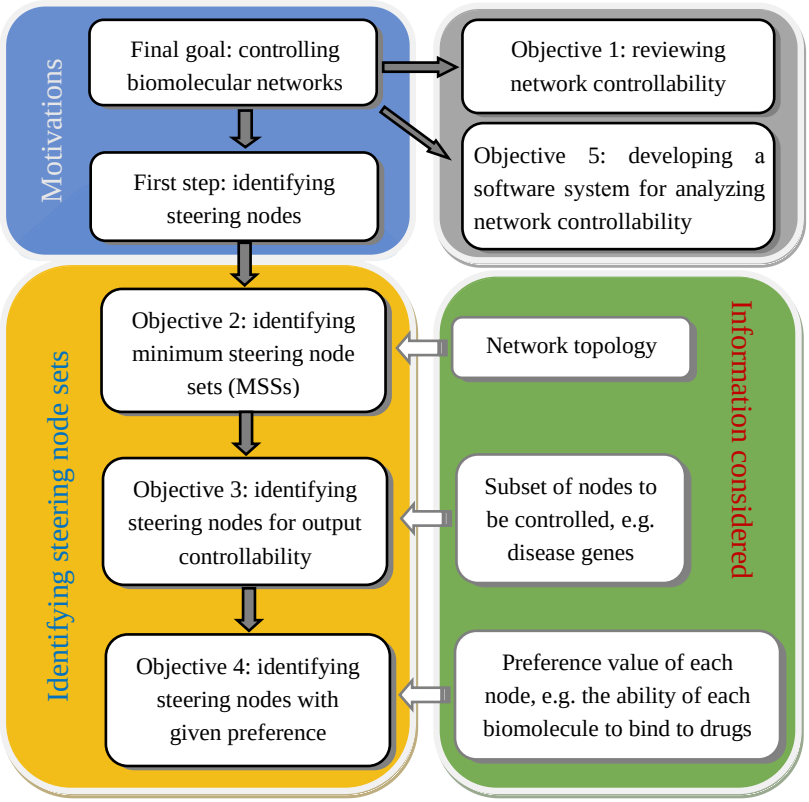


Figure 1.1: Motivation and objectives of the thesis. Objective 1 and Objective 5 aim at summarizing the studies of network controllability and realizing developed algorithms. Objective 2, Objective 3 and Objective 4 target identifying steering nodes for controlling biomolecular networks. The more information is considered, the more reliable and practical the results are.

1.3 Organization of the thesis

This is a manuscript-style thesis. The main content of the thesis consists of the works to achieve the objectives, which is presented in the form of published or prepared manuscripts. At the beginning of each chapter, a brief introduction is included to describe the connection of the manuscript to the context of the thesis. In

Chapter 7, a general discussion of correlations of each manuscript is provided. All manuscripts have been re-formatted for consistency.

The remainder of the thesis is organized as follows: Chapter 2 presents a comprehensive review of controllability of complex networks and its applications to biological networks. Chapter 3 introduces a method to identify MSSs for the complete controllability of complex networks and applications to several biomolecular networks for validation of the method. Chapter 4 proposes a method to identify steering nodes for output controllability, which aims to control a subset of nodes in a network instead of completely controlling the whole network. A typical application of the method is drug target identification: steering nodes are potential drug targets for controlling disease related biomolecules. Chapter 5 further improves the feasibility of identified steering nodes for controlling biomolecular networks by selecting steering nodes with drug binding preference such that the steering nodes have more chemical-binding opportunities with drugs. Chapter 6 presents a software system which integrates nine network controllability algorithms recently developed. Finally, Chapter 7 summarizes the work in this thesis and discusses some future work of this research. The copyright permissions of the manuscripts included are in Appendix A.

2 Controllability of complex networks and its applications to biological networks

Prepared as: L. Wu, M. Li, J. Wang, and F.-X Wu, "Controllability of complex networks and its applications to biological networks," *Journal of Computer Science and Technology*, submitted, 2018.

This chapter presents a comprehensive review of studies on the controllability of complex networks and its applications to biological networks. It first introduces different dynamic models of biological networks and compares the methods for investing controllability of each model. Then, based on linear dynamic model, methods for identifying steering node sets for various control objectives are surveyed. For applications to biological networks, biological functions or importance of steering node sets as well as individual nodes playing different roles in the controllability of biological networks are reviewed. Objective 1 of the thesis is fulfilled by this chapter.

Abstract

Complex networks have been used to represent many systems of interests in practice. Instead of performing functions alone, biological elements, such as biomolecules in biomolecular networks, neurons in neuronal networks or ROIs (region of interest) in brain networks, usually exert their functions through interactions with others to form various types of biological networks. Since biological networks are dynamic and complex, their behaviours are hard to be predicted from individual biological elements. Therefore, computational tools are needed to reveal essential biological mechanisms from a systematic perspective. Controllability, which is a concept in control theory, has been applied to investigate the dynamics of complex networks recently. Advances in the controllability of complex networks inspire investigations on the controllability of biological networks. Studies on controllability of biological networks show promising applications such as identifying potential drug targets or biologically important biomolecules. However, there is no comprehensive study for reviewing controllability of biological networks.

In this article, recent advances on the controllability of complex networks and applications to biological networks are reviewed. First, we briefly compare three dynamic models of biological networks and controllability of each model. Then we focus on complete controllability and controllable subspaces of networks based on the structural linear dynamic model, which explores the underlying correlations between the network

topology and dynamic properties. For complete controllability and controllable subspaces, algorithms for identifying steering node sets for each control objective are introduced, respectively, where steering nodes are the nodes which should be directly actuated by input control signals to achieve control objectives. Finally, we review applications of network controllability theorems to biological networks, in which biological functions and network topology have been connected from the aspect of control theory.

2.1 Introduction

Biological systems are composed of biological elements that can interact with each other. The structure of biological systems can be described by biological networks in which nodes are biological elements and edges connect biological elements that have interactions. Biological processes, which are vital for living organisms to live, are usually carried out by complicated interactions among a variety of biological elements. Therefore, studying biological elements and their interactions are critical for understanding the roles of biomolecules within cells and uncovering the mechanisms of biological processes. With the development of biomedical techniques, such as high throughput technologies and MRI (magnetic resonance imaging), various types of biological data have been acquired in a large amount, which benefits the reconstructions of different types of biological networks [3, 5, 29–32].

Network science is an interdisciplinary academic field that analyzes complex relationships between objects. Current advances in network science have shown that most real systems share numerous non-trivial topological features [17, 18, 33], which are the results of the common dynamical principles that govern their emergence and growth. By studying topological features, network science has provided powerful tools for investigation of biological networks to excavate underlying relationships among biological elements and to reveal the essential biological mechanisms from a system perspective. Substantial progresses have been made based on investigations of biological networks, such as drug target identification [6, 7], human disease gene prediction [8] and protein complex identification [9].

The ultimate goal of investigating a network is to control its behaviour or state. For biological networks, the ability of controlling their behaviours manifests the capability of changing phenotypes of biological systems as desired, which is vital for improving human lives. However, network science mainly focuses on static topological features while it does not capture dynamics of individual nodes in complex networks. Control theory, on the other hand, is a relatively well established subject in engineering which deals with the control of dynamic systems. In 1960s, Kalman pioneered the state-space approach to systems and introduced the notions of controllability and observability [34], which have become the bases of modern control theory. In the state-space representation, individual nodes have their own state variables which have specific physical meanings (e.g. gene expression levels in gene regulatory networks). Because of the interactions among nodes in a network, actuating the states of some nodes can affect other nodes, which may change the state of the network. Understanding controllability, which measures the ability to steer a system from any initial

state to any final state, is critical in implementation of controlling networks. Therefore, several fundamental questions are raised naturally: How to model the dynamics of complex networks? To what extent are the interactions among nodes or the topological features related to the dynamics and controllability of complex networks? How to select a set of steering nodes such that by perturbing the states of these nodes a network can be steered to a desired state? How to design optimal control strategies under realistic constraints? What information can we get from biological networks based on network controllability? To answer these questions, we review recent studies on the controllability of complex networks. In addition, applications of network controllability to biological networks are discussed specifically, though the developed network control methods can be applied to other complex networks.

Fig. 2.1 shows the topics and contents of following sections, which are based on the flow for analyzing controllability of biological networks. To explore the controllability biological networks, the first step is to determine the dynamic model to represent the dynamics of biological networks. In Section 2, we compare different dynamic models of biological networks. After the model comparison, we focus on the controllability of biological networks based on linear dynamic model. Section 3 and Section 4 introduce the network controllability theorems and proposed methods to identify steering node sets for different control objectives, respectively. In section 5, recent studies on the controllability of biological networks are reviewed.

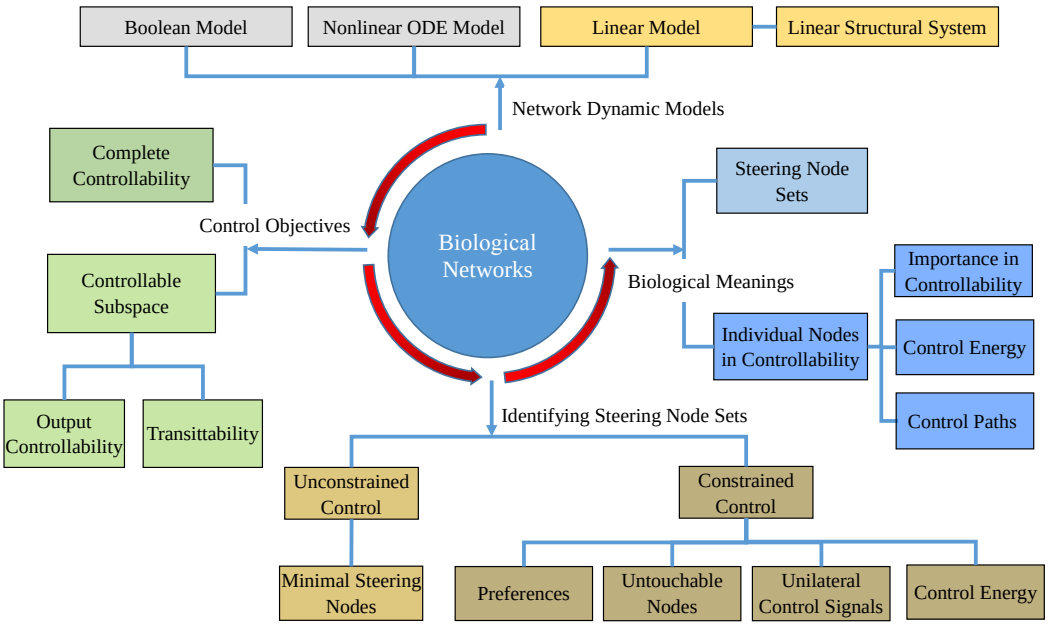


Figure 2.1: Contents in the article and flow of analyzing controllability of biological networks. For the dynamic models, we first introduce three dynamic models for representation of biological networks. Current progresses about controllability of networks with different dynamic models are discussed, respectively. In addition, we illustrate the reasons of focusing on linear dynamic model in this article. Then we introduce control theorems related to the complete controllability and controllable subspace of complex networks. Based on the control theorems, methods for identifying steering node sets for different control objectives are discussed. Finally, we review studies of biological networks from the perspective of network controllability. Biological meanings of specific steering node sets and individual nodes that play different roles in network controllability are discussed.

2.2 Dynamic models of biological networks

To understand the controllability of biological networks, it is important to make clear the dynamic models. Different dynamic models lead to different analysis methods for controlling complex networks. In this section, we present several commonly used dynamic models for biological networks and discuss the control of networks based on these models. By comparing different models, we focus on the controllability of complex networks based on linear dynamic model.

2.2.1 Boolean dynamic model

For many artificial and natural systems, state variables of individual components have two distinct configurations. In 1969, Kauffman introduced the Boolean dynamic model for gene regulatory networks [35]. In a gene regulatory network, each gene is represented by a node and the values of its state variable is either 0 or 1, which corresponds to expressed or unexpressed. Currently, Boolean dynamic model has become a powerful tool for describing and analyzing a variety of biological networks, which benefit our understanding of many biological processes [36, 37].

The Boolean dynamic model of a network with n nodes is represented by the equation:

$$\mathbf{x}(t+1) = f(\mathbf{x}(t)), \tag{2.1}$$

where $\mathbf{x}(t) = (x_1(t), ..., x_n(t))^T \in B^n$ is a state vector that describes the states of all nodes in the network. $B = \{0, 1\}$ is defined as a set that contains two Boolean values. f is a Boolean function that maps $B^n \to B^n$.

To study the controllability of Boolean networks, Cheng *et al.* developed a method to map a Boolean dynamic model into a standard discrete-time linear dynamic model, which provides an algebraic framework for investigating Boolean networks [38]. However, the idea underlying this approach is mapping 2^n possible states of a network into a $2^n \times 2^n$ matrix [39]. Therefore, it is computationally intractable to test the controllability of large-scale Boolean networks. In fact, it can be proved that finding an optimum strategy to control a Boolean network to a desired final state is an NP-hard problem [40]. Though it is computational intractable for general Boolean networks, Akutsu *et al.* studied some special cases of networks which have tree structure or contain no more than one directed cycle [41]. Kim *et al.* utilized the genetic algorithm to identify a minimum set of nodes such that by pinning the state of each node in the set to the corresponding desired final state, the whole network will eventually converge to the desired state [42]. However, genetic algorithm is a heuristic algorithm which does not guarantee the minimality of identified set.

2.2.2 Nonlinear ODE model

Since biological processes are nonlinear, modeling the dynamics of biological networks by nonlinear ordinary differential equations (ODEs) is a straightforward idea. Michaelis-Menten kinetics is one of the best-known models for describing the rate of enzymatic reactions in biochemistry. Because many regulatory processes

between biological elements are carried out by biochemical reactions, the Michaelis-Menten kinetics is suitable for modelling different types of biological networks, such as signaling networks [43], metabolic networks [12] and gene regulatory networks [44].

The general form of dynamics of a network with n nodes is represented by the following equation:

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t)), \tag{2.2}$$

where $\mathbf{x}(t) = (x_1(t), ..., x_n(t))^T \in R^n$ is an n-dimensional vector that describes the states of all the n nodes in the network and $\mathbf{u}(t)$ is an m-dimensional vector that corresponds to m independent input control signals. f is a nonlinear function.

Though great efforts have been devoted to understanding the controllability of nonlinear systems [45–47], we still lack of general methods to test the controllability of nonlinear systems. In fact, because nonlinear differential equations rarely have closed form solutions, it is not feasible to develop general theory on nonlinear controllability. Therefore, a weaker form of controllability called local accessibility of nonlinear dynamics systems has been investigated [48], where local accessibility measures the ability to reach an open set of states in the state space from a given initial state.

To steer a nonlinear dynamic network to a desired state, Cornelius [41] proposed a strategy to perturb the states of nodes such that the network can be steered into the "basin of attraction" of the desired final state [49]. Once arriving in the basin, the network will evolve spontaneously to the desired final state.

2.2.3 Linear model

Although dynamics of biological systems are nonlinear, linear models have also been applied to describe the dynamics of biological networks such as gene regulatory networks [50]. To study the controllability of biological networks, it is reasonable to represent the dynamics of biological networks by linear dynamic models. First, there are a large amount of tools available from control theory to study systems with linear dynamics. For example, a sufficient and necessary condition for the controllability of general linear systems has been developed by Kalman [34]. Second, the controllability of nonlinear systems is structurally similar to that of linear systems in many aspects. If a network is structurally controllable, then it is controllable for almost all possible parameter realizations [20]. Therefore, the structural controllability of linear system can provide a sufficient condition for the controllability of most nonlinear systems [19, 51]. Actually, to develop strategies for controlling nonlinear networks, the first step is to investigate the controllability of the locally linearized system [40]. Last but not the least, there is an intuitive connection between the network topology and the state transition matrix of linear dynamic model, which makes it possible to create dynamic model for large-scale biological networks based on their topology. Owing to the strong correlations between linear model and the network topology, studying on linear dynamic model can provide a vision of a previously proposed question, which is how much is the controllability of biological networks related to their topological features.

According to reasons discussed above, in the following sections of this article, we will focus on the controllability of linear systems and biological networks represented by linear dynamic models. For a linear time-invariant network with n nodes, the dynamics can be described by the equation:

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \tag{2.3}$$

where $\mathbf{x}(t) = (x_1(t), ..., x_n(t))^T \in R^n$ is an n-dimensional vector that describes the states of all n nodes in the network. A is an $n \times n$ state transition matrix, whose structure is determined by the adjacent matrix of the network, indicating the regulatory relationships between nodes in the network. Entry a_{ij} in matrix A indicates the intensity of influence from node j to node i. $\mathbf{u}(t)$ is an m-dimensional vector of m independent input control signals. The $n \times m$ matrix B is an input matrix indicating nodes which are directly actuated by input control signals. A network system described by the equation (2.3) is denoted as system (A, B).

2.2.4 Structural system and graph representation of linear systems

When modeling the dynamics of complex networks, the nonzero entries in matrix A indicate the strengths of relationships between nodes in the networks. However, in many scenarios, it is not feasible to obtain the values of nonzero entries in matrix A precisely. For example, although it is feasible to qualify whether there is a regulatory relationship between two nodes in biological networks, it is difficult to quantify the intensity of the regulation. In addition, though Kalman's controllability rank condition has been proposed to test the controllability of linear systems [34], calculating the rank of controllability matrix of a large-scale network is computationally intractable. Therefore, it is difficult to test the controllability of a network directly by Kalman's controllability theorem. To address these issues, recent studies on the controllability of complex networks are mainly based on the framework of structural system, which was proposed by Lin in 1974 [20]. In Lin's study, controllability of structural systems was studied and the sufficient and necessary condition for structural controllability of structural systems was given. Lin's result has been proved in different ways [52–55] and generalized to controllable subspaces [22–24].

When entries in matrices A and B are either fixed zero or independent free parameters, matrices A and B are called structural matrices and the corresponding system (A, B) is called a structural system. A structural system (A, B) is called completely structurally controllable if the Kalman's controllability condition can be satisfied by freely choosing the values of the independent free parameters in matrices A and B [20]. Besides completely structural controllability, structural output controllability [24] and structural transittability [56] have been studied, respectively.

The rationale of investigating the controllability of networks based on the structural system comes from two aspects. First, the structural linear dynamic model of a network can be created only based on its topology and each nonzero entry in A corresponds to an edge in the network. Therefore, for modeling biological networks, there is no need to consider the types of biological networks, the kinetic models regulating the dynamics as well as plenty of unknown parameters. Second, if a structural system (A, B) is structurally

controllable, most of its parameter realizations which are denoted as admissible systems $(\widetilde{A}, \widetilde{B})$ are controllable, where $(\widetilde{A}, \widetilde{B})$ can be obtained by assigning some specific values to the free parameters of (A, B) [20]. Therefore, if a network is structurally controllable, no matter how to choose the values of unknown regulatory strengths, the probability that the network is controllable is almost 100%, except some cases that the unknown regulatory strengths satisfy some constraints (equations). Therefore, the structural controllability analyses can provide reliable results for real networks even though their parameters are unknown.

Each structural system (A, B) can be represented by a digraph $G(A, B) = \{V, E\}$, where $V = V_A \bigcup V_U$ is a node set and E is an edge set. Nodes in $V_A = \{v_1, ..., v_n\}$ correspond to nodes in network under investigation and nodes in $V_U = \{u_1, ..., u_m\}$ correspond to the input control signals represented by $\mathbf{u}(t)$. $E = \{v_j \to v_i, u_k \to v_l | a_{ij} \neq 0,\ b_{lk} \neq 0\}$ consists of edges among nodes and edges from control signals to nodes. The subgraph of $G(A, B)$ induced by the node set V_A is denoted as $G(A)$, which is the original network without input control signals. Fig. 2.2 is an example of the system (A, B) and its graph representation.

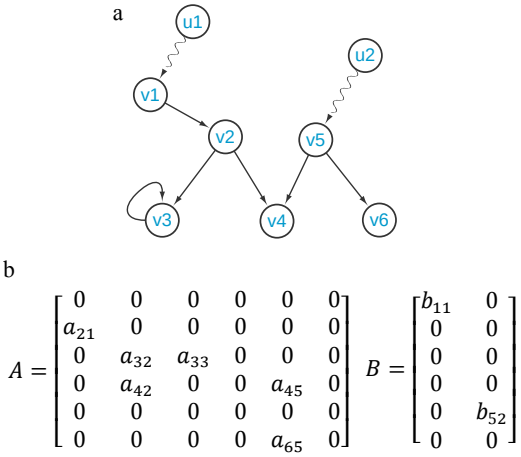


Figure 2.2: Graph representation of a network system. (a): G(A,B) corresponds to system (A, B). (b): The state transition matrix A and input matrix B of the system (A, B).

With graph representation of structural systems, algebraic structural controllability conditions can be converted to graph-theoretic forms. Therefore, various graph-theoretic algorithms can be applied to investigate structural controllability of complex networks, which is more computationally feasible for large-scale networks compared to algebraic methods. Taking the advantages of structural controllability, Liu *et al.* [19] recently applied the concept of structural controllability to complex networks. Liu's research inspired several research progresses on network controllability, such as the controllability of networks with edge dynamics [57] or nodal dynamics controllability [58], robustness of network controllability [59] and enhancing network controllability (reducing input control signals) via minimal structural perturbations [60]. In addition, a comprehensive platform has been developed for analyzing the controllability of complex networks [61].

2.3 Network controllability theorems

Since linear systems has been deeply studied in control theory, a variety of controllability theorems have been proposed, which paves the way to understand the controllability of complex networks. In addition, taking the advantages of structural controllability, connections between the network topology and controllability can be established. In this section, we introduce some important theorems of network controllability.

2.3.1 Complete controllability of complex networks

A network is completely controllable if it can be steered from any initial state $\mathbf{x}(0)$ to any desired final state $\mathbf{x}(t_f)$ in finite time t_f with appropriate control signals. Condition for complete controllability is given by the following theorem:

***Theorem* 2.1** (Kalman's controllability theorem [34])**.** System (A, B) is completely controllable if and only if the $n \times nm$ controllability matrix

$$\mathfrak{C} = \begin{bmatrix} B & AB & A^2B & ... & A^{n-1}B \end{bmatrix} \tag{2.4}$$

has full row rank of n.

To interpret this criterion, the equation 2.3 can be solved in the following form:

$$\mathbf{x}(t) = e^{At}\mathbf{x}(0) + \int_0^t e^{A(t-\tau)}B\mathbf{u}(\tau)d\tau. \tag{2.5}$$

On the right-hand side of equation 2.5, the first term corresponds to the state that the network will be without any control signals and the second term represents the effect of control signals on the network. $e^{A(t-\tau)}B$ can be expanded in series, which is a linear combination of the columns in controllability matrix $\mathfrak{C}$. When a network is completely controllable, the final state $\mathbf{x}(t_f)$ could be any state in the n-dimensional state space. On the one hand, if $rank(\mathfrak{C}) < n$, columns in $\mathfrak{C}$ will not contain a full basis to span the entire n-dimensional state space (See Fig. 2.3). Then there exist some final states $\mathbf{x}_{t_f}$, such that by letting $\mathbf{x}(t_f) = \mathbf{x}_{t_f}$, equation 2.5 has no solution for $\mathbf{u}$. On the other hand, if $rank(\mathfrak{C}) = n$, columns in $\mathfrak{C}$ contain a full basis. Given any desired final state $\mathbf{x}_{t_f}$ and let $\mathbf{x}(t_f) = \mathbf{x}_{t_f}$, an appropriate input vector $\mathbf{u}$ can always be solved based on equation 2.5. Therefore, the system is completely controllable.

For a structural system (A, B), the rank of $\mathfrak{C}$ is a function of independent free parameters in A and B. The maximum value of the rank of $\mathfrak{C}$ is defined as the *generic dimension of the controllable subspace* of structural system (A, B) and denoted by $GDCS(A, B)$. A structural system (A, B) is completely structurally controllable if and only if $GDCS(A, B) = n$, which means it is possible to choose the values of the free entries in matrices A and B such that the Kalman's controllability rank condition is satisfied.

A graph-theoretic condition for structural controllability (Theorem 2.2) has been developed in previous studies [20, 22, 23]. Before presenting Theorem 2.2, we introduce following two definitions (See Fig. 2.4 for example):

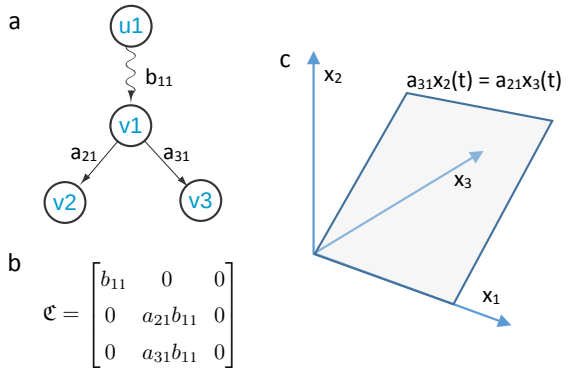


Figure 2.3: Uncontrollable network system (A, B). (a): G(A,B) corresponds to system (A, B). (b): The controllability matrix of system (A, B). (c): Controllable subspace. Suppose $\mathbf{x}(0) = \mathbf{0}$, state of the network will be kept in a subspace, which is the plane $a_{31}x_2(t) = a_{21}x_3(t)$, no matter how to choose the input control signal $u_1(t)$.

***Definition* 2.1** (Accessibility [20, 62])**.** In digraph G(A,B), a node v_i in V_A is called accessible if and only if there exists a directed path from the input vertices V_U to v_i , otherwise v_i is inaccessible.

***Definition* 2.2** (Dilation [20, 62])**.** The digraph G(A,B) contains a dilation if and only if there is a subset S of V_A such that $|T(S)| < |S|$, where $T(S) = \{v_j \ | (v_j \to v_i) \in E \ and \ v_i \in S\}$ and E is the edge set of G(A,B). The input nodes are not allowed to belong to S but belong to T(S). $|S|$ or $|T(S)|$ is the cardinality of set S or T(S), respectively.

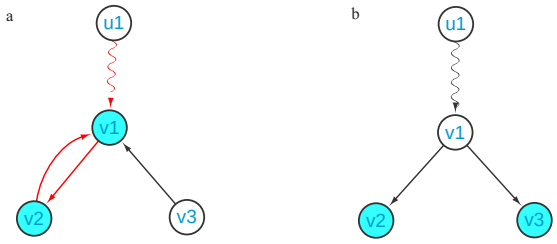


Figure 2.4: Inaccessible nodes and dilation. (a): There is no path from u_1 to v_3. Therefore, nodes v_3 is inaccessible. Nodes v_1 and v_2 are accessible. (b): Consider a set $S = \{v_2, v_3\}$, we have the $T(S) = \{v_1\}$. Because $|T(S)| < |S|$, there exists a dilation. According to Theorem 2.2, systems in (a) and (b) are both structurally uncontrollable.

***Theorem* 2.2** (Completely structural controllability theorem [20, 62])**.** A structural system (A, B) is completely structurally controllable if and only if:

i) there is no dilation in the digraph G(A,B).

ii) all nodes in V_A are accessible.

There is an equivalent expression of condition i): all the nodes in V_A can be covered by node disjoint simple cycles or simple paths starting from nodes in V_U. In a graph, a simple path is a sequence of edges

$\{(v_1 \to v_2), (v_2 \to v_3), ..., (v_{k-1} \to v_k)\}$ where all the nodes $\{v_1, v_2, ..., v_k\}$ are distinct. If $v_1 = v_k$ and other nodes are distinct, the sequence of edges is called A simple cycle.

2.3.2 Control in subspaces

In many practical problems, it is neither feasible nor necessary to completely control a network, which prompts researchers to develop methods for controlling parts of a network. Though a system may not be completely controllable sometimes, it remains controllable within a subspace (See example in Fig. 2.3c). Having a system controllable within a subspace is enough for many real applications. In addition, it is natural that ensuring the controllability within a restricted subspace will require fewer steering nodes being actuated by input control signals than ensuring controllability within the whole state space. Therefore, several approaches have been proposed to investigate the controllability of networks within subspaces.

Controllable subnetwork

For structural system (A, B), the dimension of its controllable subspace is measured by $GDCS(A, B)$, which is the maximum rank of the controllability matrix $\mathfrak{C}$ by arbitrarily choosing the values of independent free parameters. Hosoe has proved [22] that if all nodes in a network system (A, B) are accessible, then

$$GDCS(A, B) = \max_{G \in G^*} \{|E(G)|\}, \tag{2.6}$$

where G^* denotes the set of subnetworks of $G(A, B)$ which can be spanned by a collection of vertex-disjoint cycles and at most m simple paths (correspond to m control signals). $|E(G)|$ is the number of edges in G. Actually, each subnetwork in G^* is completely controllable. Therefore, the dimension of its controllable subspace $GDCS(A, B)$ equals to the number of edges in the largest controllable subnetwork in set G^*. Consider the structural system (A, B) in Fig. 2.3, the corresponding G^* consists of two subnetworks of $G(A, B)$ induced by node sets $\{u1, v1, v2\}$ and $\{u1, v1, v3\}$, respectively. According to Hosoe's controllable subspace theorem, the $GDCS(A, B)$ of system in Fig. 2.3 is 2, which suggests the whole network can be steered in a 2-dimensional state space. Suppose the network is at the origin at time $t = 0$, it can be observed that the states of nodes $v2$ and $v3$ must satisfy the equation $a_{31}x_2(t) = a_{21}x_3(t)$. But if we only need to control the subnetwork induced by nodes $\{v1, v2\}$ or $\{v1, v3\}$, it is enough by actuating node v_1 alone.

Based on Hosoe's controllable subspace theorem, recent studies investigated the controllable subspaces or completely controllable subnetworks from different perspectives, which supplement the theoretical foundation of structural controllability of complex networks. Liu *et al.* [63] defined the control centrality to measure the ability of individual nodes to control a network. Control centrality of node i is defined as $GDCS(A, b(i))$, where $b(i)$ is a vector with a single nonzero ith entry. The higher control centrality of node means by actuating only the node with input control signal, the whole network can be steered in a larger dimension of its state space or a larger subnetwork can be completely controlled. Control centrality can be extended to cases in which more than one node are actuated by control signals. Based on this observation, Iudice *et*

al. [64] introduced the network permeability, which measures the propensity of a network to be controllable. To calculate the permeability, Indices solved a problem related to control centrality at first: identifying m steering nodes from a network with n nodes, such that the corresponding $GDCS(A, B_m)$ is maximized, where B_m is an $n \times m$ controllability matrix corresponding to m steering nodes. Then the permeability is defined as:

$$\begin{aligned} \mu &= \frac{\int_0^n (GDCS(A, B_m) - m)dm}{\int_0^n (n - m)dm} \\ &= \frac{2}{n^2} \int_0^n (GDCS(A, B_m) - m)dm. \end{aligned} \tag{2.7}$$

According to the definition, for a network with a high permeability, a large controllable subspace can be obtained or a large subnetwork can be completely controlled by actuating a relatively small set of steering nodes. In order to find a subnetwork which is easy to be controlled with less steering nodes, Liu and Pan [65] proposed a method to choose subnetworks that are important and easy to be controlled in network systems. Then the authors applied this method to multiple real networks and discovered that nodes in the subnetworks chosen by this method tend to be essential. In another study, Commault *et al.* [66] claimed that though the dimension of the controllable subspace is constant for almost any parameter realization of a structural system (A, B), the subspace itself is a function of these parameters. Therefore, the authors defined a concept called fixed controllable subspace, which is the intersection of the controllable subspaces of all parameter realizations whose dimension of controllable subspace equals to $GDCS(A, B)$.

Output (Target) controllability

In real applications, we are interested in controlling a specific subset of nodes or a subnetwork of interest. Since the subset of nodes can be considered as the output of the network, Wu *et al.* [67] formulated the problem of controlling a predefined subset of nodes in a network as a network output controllability problem. Gao *et al.* [68] proposed the same idea in an independent study, in which they referred to as target control.

The outputs of a linear dynamic system (A, B) can be described by the following equation:

$$\mathbf{y}(t) = C\mathbf{x}(t), \tag{2.8}$$

where $\mathbf{y}(t) = (y_1(t), ..., y_p(t))^T$ is an output vector in which each entry represents an output. C is a $p \times n$ matrix that indicates the outputs of the network. A system described by equations (2.3) and (2.8) is denoted by matrix triplet (A, B, C). For target controllability, the outputs are defined as the states of a set of nodes in the network. Then it is assumed that there is one and only one nonzero entry in each row of C such that $\mathbf{y}(t)$ is a p-dimensional vector that each entry corresponds to the state of one node. Therefore, target controllability is a special case of output controllability.

A network is output controllable if its outputs can be steered from any initial state $\mathbf{y}(0)$ to any desired final state $\mathbf{y}(t_f)$ in finite time t_f with appropriate control signals. To test the output controllability of a system (A, B, C), a $p \times mn$ output controllability matrix is defined as:

$$\mathfrak{o}\mathfrak{C} = \begin{bmatrix} CB & CAB & CA^2B & ... & CA^{n-1}B \end{bmatrix}. \tag{2.9}$$

The condition of output controllability is given by the following theorem in control theory:

***Theorem* 2.3** (Output controllability theorem [69])**.** System (A, B, C) is output controllable if and only if $\text{rank}\,(\mathfrak{o}\mathfrak{C}) = p$.

For a structural system, the rank of $\mathfrak{o}\mathfrak{C}$ can reach a maximum value by arbitrarily choosing the values of independent free parameters in A, B and C. The maximum value is defined as the *generic dimension of the controllable output subspace* of structural system (A, B, C) and denoted by $GDCOS(A, B, C)$. The structural system (A, B, C) is called structurally output controllable if $GDCOS(A, B, C) = p$ [24, 70]. Though theorem 2.3 presents conditions for output controllability, there is no method to calculate $GDCOS(A, B, C)$ of structural system (A, B, C). Murota and Poljak [24] have developed a method to determine the upper and lower bounds of $GDCOS(A, B, C)$.

Transittability

Output controllability measures the ability of a predefined subset of nodes that can be steered by input control signals. However, the states of nodes out of the predefined subset are not considered during control processes. On the other hand, Wu *et al.* [56] introduced a new concept called transittability of networks, which measures the ability of transition between two specific states of complex networks. Transittability takes the states of all nodes into consideration as well as reduces the required steering nodes compared to complete controllability.

For system (A, B), it is called transittable between these two specific states $\mathbf{x}_0$ and $\mathbf{x}_1$ if there exists input control signals $\mathbf{u}(t),\ t \in [0,\ t_f]$, by which the system (A, B) can be transited between two specific states $\mathbf{x}(0) = \mathbf{x}_0$ and $\mathbf{x}(t_f) = \mathbf{x}_1$. A sufficient and necessary condition for transittability controllability is given by the following theorem:

***Theorem* 2.4** (Transittability theorem [56])**.** With either specific state $\mathbf{x}_0$ or $\mathbf{x}_1 \in span\{\mathfrak{C}\}$, system (A, B) is transittability between $\mathbf{x}_0$ and $\mathbf{x}_1$ if and only if

$$rank(\mathfrak{C}) = rank(\bar{\mathfrak{C}}),$$

where $\bar{\mathfrak{C}} = \begin{bmatrix}\bar{B} & A\bar{B} & A^2\bar{B} & \ldots & A^{n-1}\bar{B}\end{bmatrix}$ and $\bar{B} = [x_0 - x_1, B]$.

Similarly to structural controllability, a structural system (A,B) is called structurally transittable between two specific structural states $\mathbf{x}_0$ and $\mathbf{x}_1$ if there exists an admissible system $(\widetilde{A}, \widetilde{B})$ (with respect to (A, B)) and admissible states $\widetilde{\mathbf{x}_0}$ and $\widetilde{\mathbf{x}_1}$ (with respect to $\mathbf{x}_0$ and $\mathbf{x}_1$, respectively) such that the system $(\widetilde{A}, \widetilde{B})$ is transittable between states $\widetilde{\mathbf{x}_0}$ and $\widetilde{\mathbf{x}_1}$. In fact, for structure systems, the transittability between two structure states actually measures the ability to control a subset of nodes in the network without disturbing other nodes.

2.4 Identification of steering node sets

In order to control a network, the first step is to identify a set of steering nodes which should be actuated by input control signals. A network system is completely controllable if each node is directly actuated by a distinct input control signal. However, it is costly and impractical for large networks. Therefore, methods are required to identify minimal steering node sets such that the the control objective can be satisfied. The identification of steering nodes for controlling networks can be viewed as problems of determining appropriate control matrix B when a network, which is represented by A, is given. Theorem 2.2, 2.3 and 2.4 provide conditions to judge if a structural system $(A, B)/(A, B, C)$ is completely structurally controllable, structurally output controllable or transittable between two specific states. However, for a given network (matrix A), controllability theorems do not indicate a set of steering nodes (matrix B) such that the network system is controllable. A brute-force search for a minimal steering node set would require checking the controllability conditions for almost 2^n distinct controllability matrices B, which is computationally prohibited. In this section, methods for identifying steering node sets for different control objectives are reviewed.

2.4.1 Steering nodes for complete controllability

The minimum driver node set (MDS) [19] and the minimum steering node set (MSS) [62] are two mostly investigated steering node sets for completely controlling networks. Recently, graph-theoretic methods have been proposed to identify MDSs and MSSs of networks based on theorem 2.2.

The MDS is a minimum set of nodes in which each node should be actuated by an independent control signal such that the condition i ("no dilation" condition) of structural controllability theorem 2.2 can be satisfied. However, applying independent control signals to an MDS does not guarantee complete controllability of the network and it is a necessary condition for completely controlling a network. In [19], identification of MDS has been formulated as a maximum matching problem in an undirected bipartite graph corresponding to the original network. A matching on an undirected graph is a set of edges without common nodes and a maximum matching is a matching with the largest size. To identify an MDS, a bipartite graph which contains node sets $R = \{r_1, \cdots, r_n\}$ and $C = \{c_1, \cdots, c_n\}$ is constructed. The nodes r_i and c_i correspond to the node i of G(A). If there is a directed edge from node i to j in G(A), there is an edge in the bipartite graph connecting r_j and c_i. A maximum matching in bipartite graph can be solved by Hoproft-Karp algorithm [71]. Then the MDS are corresponding to the nodes in R that are not connected to any matching edges (See Fig. 2.5b). It can be verified that if each node in MDS is actuated by an input control signal, which means adding a control node u_i for each node i in the MDS, the resulting graph $G(A, B)$ will have no dilation. Since MDSs of a network are not unique, Zhang *et al.* [72] proposed a preferential matching algorithm to identify MDSs that have a specific degree property.

The MSS is a minimum set of nodes in a network which should be actuated by control signals to completely structurally control the network. Compared to the MDS, applying independent control signals to an MSS

guarantees the complete controllability of the network and it is a sufficient and necessary condition for completely controlling a network, which satisfies both conditions of structural controllability theorem 2.2. In [62], the bipartite graph in identification of MDS has been extended to a directed graph. The authors proved that a minimum cost maximum flow (MCMF) in the constructed digraph corresponds to an MSS of the network (See Fig. 2.5c). The algorithm for solving the MCMF problem can be found in [73]. Similar to the MDS, MSSs of a network are not unique as well. Therefore, Wu *et al.* [74] developed an approach to identify MSSs with preference, such that the average preference value of nodes in the identified MSS is the maximum among all possible MSSs of the network.

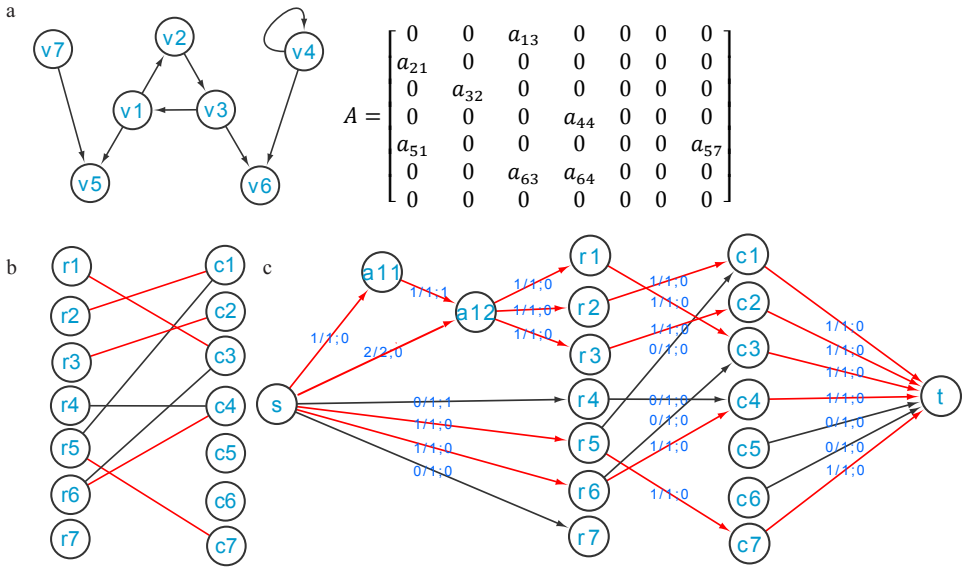


Figure 2.5: Identification of an MDS and an MSS by maximum matching and minimum cost maximum flow method, respectively. (a): A network G(A) and its corresponding system state transition matrix. (b): The corresponding undirected bipartite graph and the maximum matching. Nodes r_4 and r_7 in node sets R are not matched in the maximum matching, which suggests v_4 and v_7 make up an MDS. (c): A directed graph constructed according to the structure of G(A) and the minimum cost maximum flow. The labels on edges represent flow, capacity and cost per unit flow, respectively. There is no flow passing through r_4 and r_7, which suggests v_4 and v_7 belong to an MSS. The flow passes through one edge with cost 1, which means there is an additional steering node which should be chosen from the corresponding source strong connective component (SCC). Then nodes v_4, v_7 and $v_i (i = 1, 2, 3)$ make up an MSS.

Several studies investigated the identification of steering nodes under constraints, which are common in real applications. Pequito *et al.* [75] proved that the minimum constrained input selection (minCIS) problem, which selects minimum number of inputs from a given set of possible inputs, is NP-hard. When there are n possible inputs and each input can actuate a distinct node of the network, the minCIS problem reduces to the problem of identification of MSS, which could be solved in polynomial time.

Stepping out of structural controllability, some studies considered constraints from the aspects of input control signals and control energy. Lindmark and Altafini [76] studied the controllability of complex networks

with unilateral inputs, which assumes that an input control signal is either negative or positive, but not both. This constraint on control signals makes sense in many scenarios. For instance, input control signals of biological networks are usually drugs or chemical molecules, which can only either activate or inhibit their targets nodes. Conditions for unilateral controllability have been formulated algebraically in terms of eigenspaces of the system matrix A. Compared to unconstrained control, more steering nodes are required to achieve complete controllability for unilateral control. By studying several instances with randomly weights assigned to the edges, the authors discovered that the number of additional steering nodes for unilateral control is strongly related to the number of roots and dilations in a network. Then a lower bound of minimum number of steering nodes required for unilateral controllability can be determined by network structure alone.

In many cases, though actuating an MSS can completely control a network theoretically, the associated control cost can be unbearably large, which prevents actual control from being realized physically. The control cost can be measured by control energy, which is defined as:

$$E(t_f) = \int_0^{t_f} (\mathbf{u}_t^T \cdot \mathbf{u}_t) dt, \tag{2.10}$$

where $\mathbf{u}_t$ are input control signals [77]. Wang *et al.* [78] proposed physical controllability which considers the probability of achieving control practically. By investigating control energy for controlling chain structures in networks, the authors provided strategies to make physically uncontrollable networks physically controllable by properly adding additional steering nodes. Li *et al.* [79] studied the problem of identifying a fixed number of steering nodes, such that a network can be completely controllable with the minimum energy. The authors formulated the original problem as an optimization problem and developed two methods to solve it.

2.4.2 Steering nodes for output controllability

It has been proved that identifying the minimum number of steering nodes for structural output controllability is an NP-hard problem [80], where the outputs are defined as the states of a set of nodes. In Wu's study [70], the lower bound of $GDCOS(A, B, C)$ [24] has been applied to design an algorithm to identify steering nodes for output controllability, which guarantees that the network is output controllable by actuating identified steering nodes. Therefore, actuating the identified steering node set is a sufficient but may not be a necessary condition for structural output controllability. The identification of steering nodes for output controllability has been formulated to maximum weight complete matching problem in a bipartite graph constructed according to network topology and a predefined set of nodes to be controlled. The maximum weight complete matching problem can be solved by the Kuhn-Munkres (KM) algorithm [81]. Fig. 2.6 is an illustrative example for identifying steering nodes for controlling a subset of nodes in a network.

In Gao's study [68], a greedy algorithm has been developed to identify steering nodes which are sufficient for target control. Several further studies developed algorithms to identify steering nodes for target controllability by reducing the number of steering nodes or considering realistic constraints. Instead of using the greedy algorithm, Zhang *et al.* [82] developed an algorithm which elaborately rearranges the matching order

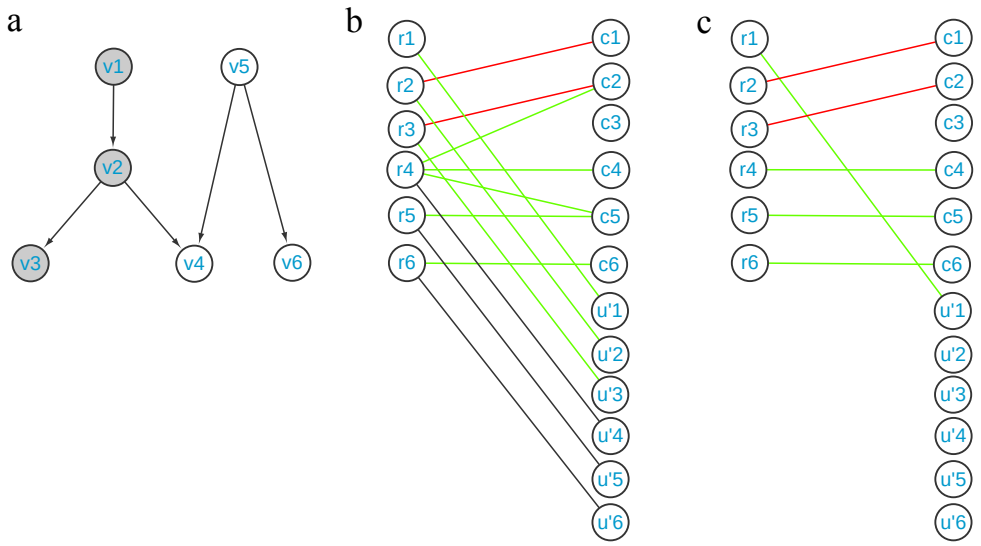


Figure 2.6: Identifying steering nodes for output controllability. (a): A network. The outputs of the system are the states of grey nodes v_1, v_2 and v_3. (b): A corresponding weighted bipartite graph. The weights of red lines, green lines and black lines are 1, 0 and -1, respectively. (c): A maximum weight complete matching in (b). Since r_1 matches a node in set $U' = \{u'_i | i = 1, \ldots, 6\}$, node v_1 makes up a steering node set for output controllability, which suggests that the states of nodes v_1, v_2 and v_3 can be controlled by actuating node v_1 only.

of the nodes such that the required number of steering nodes for target control can be significantly reduced. The comparison results on model generated networks and real networks indicate that the proposed algorithm outperforms Gao's algorithm [68]. Because the functions of network systems intensively depend on the connections between nodes, Liu *et al.* [83] investigated target controllability of giant connected components of directed networks by selecting target nodes from giant connected components, which are the connected components of networks that have constant fractions of nodes in networks. In the study, the relationships between the number of steering nodes for controlling giant connected components and the parameters of model generated networks are explored. Piao *et al.* [84] considered controlling a subnetwork called target community of a complex network when the whole topological structure of the network is not available. The authors argued that though a target community is controllable with steering nodes identified by structural controllability analysis, determining input control signals that are able to achieve a given control goal can be very difficult. It is because the process of controlling target communities would be influenced by signals from the remainder network, but the topology and state of the remainder network is not available. To deal with this issue, the author defined a type of steering nodes, which they refer to as immune nodes, for blocking signals transmitting from the remainder network. Then they proposed a method to reduce the total number of steering nodes and immune nodes such that the subnetwork is completely controllable and the signals from the reminder network can be blocked.

By considering some practical constraints, Guo *et al.* [85] proposed the concept called constrained target controllability of complex networks, which concerns the target controllability by selecting steering nodes from

a predefined constrained node set. Then the authors developed an algorithm to identify steering nodes from a constrained node set for controlling a set of target nodes. Iudice *et al.* [64] also investigated the target controllability of networks by not only considering the constraints on selection of steering nodes, but also introduced a set of untouchable nodes, whose states should not be perturbed during the control process.

2.4.3 Steering nodes for transittability

In addition, Wu *et al.* [56] developed an algorithm to identify steering nodes with a given network $G(A)$ and a set of nodes whose states are supposed to be changed during state transition. Identification of steering nodes for transittability has been formulated to maximum weight complete matching problem in a bipartite graph constructed according to network $G(A)$ and structural states. Fig. 2.7 is an illustrative example of identifying steering nodes for state transittability. The result indicates that by actuating nodes v_1 and v_3 with input control signals, the states of nodes v_1, v_2 and v_3 can be controlled without affecting nodes v_4, v_5 and v_6, which is different from the example of output controllability in Fig. 2.6: the states of nodes v_1, v_2 and v_3 can be controlled by actuating v_1; however, the state of v_4 might be perturbed as well. Transittability usually needs more steering nodes for controlling the states of target nodes compared to output controllability, but requires less steering nodes than completely controlling the whole network.

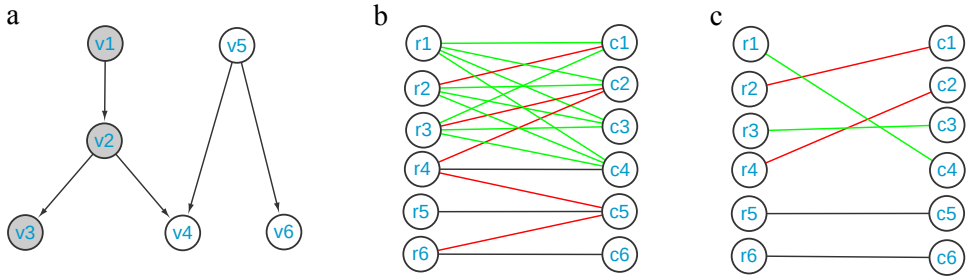


Figure 2.7: Identifying steering nodes for state transittability. (a): A network. Grey nodes v_1, v_2 and v_3 could be changed to any states in finite time while states of white nodes would not be affected by control signals at the end of the control process. (b): A corresponding weighted bipartite graph. The weights of red lines, green lines and black lines are 1, λ and 0, respectively, where λ is a small enough positive number. (c): A maximum weight complete matching in (b). Since r_1 and r_3 match green lines, nodes v_1 and v_3 make up a steering node set for state transittability.

2.5 Applications to biological networks

Studies introduced in previous section offer powerful tools to systematically identify steering node sets in controlling states of complex networks. The steering nodes are identified based on the control theory, which guarantees the controllability of networks theoretically. However, theoretical analyses on structural controllability of complex networks do not explain how to manipulate the steering nodes to steer networks from one state to another, which depends on the details of connections and interactions in the networks. Therefore, there is a gap between "controllable in principle" and "controllable in practice" [27]. Though structural

controllability approaches are not able to provide explicit strategies to control complex networks in practice, they still offer a novel perspective to investigate topology of complex networks from the aspect of control theory. In this section, we review recent studies that explored various types of biological networks, including biomolecular networks, neuronal networks and brain networks, based on network controllability.

2.5.1 Steering node sets in biological networks

By applying input control signals to steering node sets, networks can be steered to the desired states. Therefore, recent studies explored the biological meanings of steering node sets in biological networks for different control objectives, which are complete controllability, output controllability and transittability.

MDSs of biological networks are one of the most investigated steering node sets for completely controlling networks. Khazanchi *et al.* [86] compared driver nodes in MDSs and hub nodes of 4 different protein-protein interaction (PPI) networks. They found that hub nodes are more likely to be lethal proteins while driver nodes tend to be transcription factors. In addition, they found that driver nodes are enriched in first-degree neighbors of hubs, which suggests that one should control the nodes interacting with the hubs, instead of controlling the hubs directly, to control networks. Badhwar and Bagler [87] identified the MDS of *C. elegans* neuronal network. By investigating the phenotypic properties and the genetic correlations of driver neurons in the MDS, they found that driver neurons are primarily motor neurons located in the ventral nerve cord and contribute to biological reproduction, which demonstrates the importance of driver neurons and their ability of controlling the behaviours of the organism. Noori *et al.* [88] constructed a comprehensive neurochemical network of the rat brain and identified an MDS of the rat brain network. Interestingly, one of the four steering nodes in the identified MDS, subthalamic nucleus (STh), has already been proved to be crucial in global circuit dynamics [89] and treatment of Parkinson disease as well as other disorders [90] by numerous deep brain stimulation studies. This observation manifests an agreement between structural controllability and function of neuronal networks.

For the MSSs of biological networks, Wu *et al.* [62] applied their method of identifying MSS to the *S.cerevisiae* cell cycle networks [11, 13], Epithelial to Mesenchymal Transition (EMT) network [91] and myeloid differentiation regulatory network [37]. It has been discovered that steering nodes in MSSs of these networks play critical roles in triggering cell division process, maintaining homeostasis of epithelial or regulating early myeloid development as well as hematopoietic stem cells, respectively. Since the identified MSSs in these networks are closely related to dynamic behaviours of the networks, it is fair to suggest their importance in controlling the networks. In a further study on MSSs, Wu *et al.* [92] improved the algorithm of identifying MSSs by considering the preference of individual nodes, such that nodes in the identified MSS have higher preference values compared to nodes in other eligible MSSs of a network. The algorithm has been applied to study MSSs with drug binding preference of some biological networks. The biomolecules in the MSSs with binding preference are enriched with known drug targets and are likely to have more chemical-binding opportunities with existing drugs compared with randomly chosen MSSs, suggesting novel applications for

drug target identification and drug repositioning.

There is an intuitive application of output controllability to biological networks, which is drug target identification. Wu *et al.* [70] formulated the problem of drug target identification as a problem of identifying steering node set for output controllability of biological networks. In the study, disease biomolecules and biomolecules whose state changes would lead to side effects are defined as the outputs of the network, which takes both efficiency and safety into consideration. The steering nodes for controlling these two types of biomolecules are considered as potential drug targets. The method has been applied to several real biological networks. The identified potential drug targets are targets of approved drugs or in agreement with existing research results, which indicates the feasibility of the method. By considering the constrained target controllability, Guo *et al.* [85] applied the developed algorithm for identifying steering nodes to a gene regulatory network related to type 1 diabetes. By defining the five genes related to type 1 diabetes as the target nodes and all FDA-approved drug targets as the constrained node sets, they found that FASLG and CD80 are steering nodes for controlling the target nodes related to type 1 diabetes, which is supported by previous wet experiments. In another study, Kanhaiya *et al.* [93] built three PPI networks for breast, pancreatic and ovarian cancer, respectively. The authors considered survivability-essential proteins specific to each cancer type as control targets in each network. A method is also proposed to identify steering nodes among FDA-approved drug target nodes. Different from the method in Guo's study [85], selection of steering nodes from FDA-approved drug targets is not obligatory, but is preferred in Kanhaiya's study. The results indicate that many steering nodes are known drug targets for cancer therapies, but some of them are not the drug targets corresponding to cancer types. Besides identifying steering nodes for drug target identification, Yan *et al.* [94] predicted the involvement of each *C. elegans* neuron in locomotor behaviours by formalizing the responsive mechanism of *C. elegans* to external stimuli as a target control problem. The predictions based on the target control of network have been validated by their experiments. For example, it has been predicted that three neurons (DD04, DD05, or DD06) in the DD motor neuronal class should affect locomotion when ablated individually. Their experimental validation shows that ablations of DD04 or DD05 have impacts specifically on posterior body movements, whereas ablations of other neurons in the DD motor neuronal class (DD02 or DD03) do not affect the locomotion. Yan's study not only provides a novel method to unveil how the structure of neuron network affects its functions based on controllability perspective, but also offers the first experimental proof of the validity of network structural controllability analyses.

For transittability, Wu *et al.* [56] employed different biological systems with different phenotypes to validate the applicability of identified steering node sets for transittability. For example, T helper cells (Th cells), which play an important role in the immune system, are a sub-group of lymphocytes. A network has been constructed by Mendoza [95] to model the differentiation of Th cells. Matured Th cells can be classified as Th0 (precursor), Th1 and Th2 (effector) cells, which correspond to three different states of Th differentiation network. Steering node sets for state transitions among these three phenotypes have been identified by the proposed algorithm. According to the transittability analyses, actuating steering nodes

SOCS1 and T-bet can steer the network between Th0 and Th1 and actuating nodes IL-4 and GATA3 can steer the network between Th0 and Th2, which is in agreement with existing knowledge [96, 97]. Actuating steering nodes T-bet and GATA3 can cause the transition between Th1 and Th2, which is completely in agreement with the experimental data [98].

2.5.2 Roles of individual nodes in controllability

Instead of focusing on specific steering node sets for different control objectives, several studies proposed methods to quantify the importance or analyze the roles of individual nodes in controlling networks and then investigated biological meanings of nodes based on the proposed methods. The analyses were mainly based on the importance of nodes in network controllability, control energy and control paths.

Since the MDSs or MSSs of a network are not unique, the algorithms for identifying MSSs or MDSs do not result in a unique set of MDS or MSS. Therefore, studies attempted to figure out the importance of nodes in network controllability by classifying nodes into different categories or assigning centrality values to individual nodes according to the times of a node appears in all possible MDSs of a given network. Jia *et al.* [99] classified a node in the network as critical, intermittent or redundant if it acts as a driver node in all, some or none of all possible MDSs, respectively. By classifying nodes in a human signaling network, Liu *et al.* [100] discovered that critical nodes are enriched in the group of ligands, intermittent nodes are enriched in cell surface receptors and redundant nodes are enriched in intracellular signaling proteins. They also found that cancer-associated genes are enriched in redundant nodes, which suggests that controlling the regulators of the cancer-associated genes could be more feasible than controlling the cancer-associated genes directly. In a related work of the classification, Jia *et al.* [101] proposed a concept called control capacity, which is defined as the likelihood that a node is a driver node in an arbitrary MDS. Liu *et al.* [102] calculated the control capacity of nodes in a human liver metabolic network and classified nodes into critical, high-frequency and low-frequency nodes based on their control capacity values. They found that in the metabolic network, critical metabolites are likely to be essential metabolites while the high-frequency metabolites tend to participate in different metabolic pathways.

Though the MDSs of a network may not be unique, the cardinality of all the MDSs are the same. Vinayagam *et al.* [103] classified a node in a network as indispensable, neutral or dispensable, which correlated to increasing, no effect, or decreasing the cardinality of the MDSs of the network by removing that node and edges which connected to the node. Then the authors applied their classification strategy to a directed human PPI network and found that indispensable proteins or corresponding genes are enriched in essential genes, human virus targets, drug targets or disease-causing mutations. Their study provides a novel classification strategy based on network controllability. Nodes in different categories show distinct biological properties in the context of essentiality, evolutionary conservation, and regulation of translational or post-translational modifications. In fact, before the work of Vinayagam *et al.*, Matsuoka *et al.* [104] identified the indispensable nodes, which they called "critical node" in their study, of an influenza A virus life cycle network. They found

that the indispensable nodes are important factors of the viral life cycle, which are known drug targets or could be potential therapeutic targets. In another work, Uhart *et al.* [105] studied a directed phosphorylation-based PPI network by analyzing the biological characteristics of indispensable nodes. Because post-translational modification and inhibition of transduction by miRNAs are two important mechanisms of regulation in eukaryotic cells, it is meaningful to evaluate the relationship between proteins that are important in controlling the network and these two mechanisms. It has been discovered that indispensable nodes are more enriched in post-translational modifications and miRNA targets, which indicates that indispensable nodes are targets of intense biological regulation. Uhart's study provides a deeper understanding of the controllability of biological networks and bridges the controllability theorems and cell regulation processes, such as post-translational modification, in a phosphorylation-based PPI network.

In a recent work, Ravindran *et al.* [106] combined two types of classification strategies and investigated a cancer signaling network. Nodes are classified as critical, intermittent or redundant based on Jia's classification strategy and indispensable, neutral or dispensable based on Vinayagam's classification strategy. Then authors analyzed the distribution of cancer genes and targets of anti-cancer drugs in each node class. Enrichment analyses show that redundant nodes, especially indispensable redundant nodes are enriched in both cancer genes and anti-cancer drug targets, which implies a strong correlation between indispensable redundant nodes and cancer development or cancer treatment. This study indicates that the two classification strategies can capture the roles that individual nodes play in controlling a network from different aspects. Therefore, it is likely to get more comprehensive results by combining these two classification strategies.

By investigating topological features of steering nodes in MDSs, Ruth *et al.* [107] found that each driver node in MDSs corresponds to one of three topological features: source nodes, external dilations and internal dilations. Sources nodes are nodes that have no incoming edges and the number of source nodes is denoted as N_{source}. External dilations appear when sink nodes, which are nodes without outgoing edges, outnumber source nodes. The number of sink nodes is denoted as N_{sink} and then the number of external dilations equals to $N_{external} = max(0, N_{sink} - N_{source})$. Internal dilations are dilations other than external dilations and the number of internal dilation is denoted as $N_{internal}$. Then the cardinality of MDSs $N_{MDS} = N_{source} + N_{external} + N_{internal}$, which is the sum of the three topological features. Then the driver nodes can be classified into three categories based on their corresponding topological features. The authors discovered that the MDSs of a network is usually dominated by a specific topological feature. According to the proportions of each types of driver nodes, a network can be classified as source dominated, external-dilation dominated or internal-dilation dominated. The classification of networks has been tested on various of real networks. The results offer insights into the relationship between topology and functions of complex networks. For example, neural networks are source dominated, which tend to allow relatively uncorrelated behaviors and are suitable for distributed processing.

Control energy has also been applied to uncover the roles of individual nodes in the controllability of networks. Gu *et al.* [32] studied the controllability of a human brain network, in which each node represents

a region of interest (ROI) of the human brain. Three types of measures are developed to quantify the importance of nodes in controlling the brain network: average controllability measures the ability of brain regions to steer the system state with less energy input; modal controllability identifies brain regions that steer the system to states which require substantial input energy and boundary controllability identifies brain regions that locate at boundaries between network communities and control the segregation and integration of cognitive systems. This study provides a novel perspective to understand the cognitive processes from the control energy in network control. The proposed methods and measures based on control energy could provide insights into studies on the controllability of other types of biological networks.

To understand disease etiology from the perspective of network control, Wang *et al.* [108] defined a concept called perturbation influence, which is a subset of nodes based on the control paths (vertex-disjoint cycles and simple paths starting from steering nodes), to identify and quantify the ways by which disease genes perturb human regulatory networks. Intuitively, for a certain disease, the perturbation influences of different disease genes can be considered as the significant pathways related to the disease, which are etiologically essential. In addition, perturbation influence can be applied to prioritize disease genes according to the similarities of perturbation influences between nodes and known disease genes. Validation of the prioritizing method on 112 diseases shows that this method outperforms the state-of-art method PRINCE [109]. Similar to perturbation influence, concepts such as control range [110] or vertex domination centrality [111], which are defined based on control paths as well, have been proposed to study the controllable subspaces of nodes or measure the importance of nodes in controlling networks. The proposed controllability concepts based on control paths enrich analytical tools for understanding roles of nodes in controlling network subspaces.

2.6 Conclusion and discussion

In this article, we have reviewed recent advances on the controllability of complex networks and the applications to biological networks. First, different dynamic models of complex networks were briefly reviewed. Because of the effectiveness and practicability, we focused on studies which explore the controllability of complex networks based on linear dynamic model. Then we reviewed algorithms to identify steering node sets for complete controllability or controlling subspaces. Biological meanings of nodes which play different roles in controlling biological networks were investigated.

Besides controllability, other concepts in control theory could also shed lights on our ability to understand or manipulate biological networks, which is worthy for future investigation. For example, observability, which is a mathematical dual problem of controllability, can be applied to measure the states of biological networks by monitoring a specific set of biological elements. Other practical constraints such as control trajectories can be considered in order to avoid some forbidden or fatal states of biological networks during control processes. It is believed that controlling biological networks will be increasingly feasible and effective when our knowledge of control theory is enhanced and our understanding of dynamics of biology systems is deepened.

Acknowledgments

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the National Natural Science Foundation of China under Grant No. 61772552 and No.61622213, and Chinese Scholarship Council (CSC).

3 Minimum steering node set of complex networks and its applications to biomolecular networks

Published as: L. Wu, M. Li, J. Wang, and F.-X. Wu, "Minimum steering node set of complex networks and its applications to biomolecular networks," *IET Systems Biology*, vol. 10, no. 3, pp. 116-123, 2016.

In the previous chapter, we have reviewed diverse investigations of biomolecular networks from the aspect of network controllability. Most methods developed to explore the controllability of biomolecular networks are based on a specific type of steering node set — MDS. However, from the perspective of control theory, applying independent input control signals to an MDS only satisfies one of two conditions for completely controlling a network, which does not guarantee the complete controllability of the network. Therefore, it is reasonable to assume that identifying a more accurate steering node set, which satisfies both conditions of complete controllability, will benefit our understanding of controllability of biomolecular networks.

In this chapter, a graph-theoretic based algorithm is proposed to identify the MSS of a network while applying independent input control signals to an MSS is a sufficient and necessary condition to ensure the complete controllability of the network. Applications to biomolecular networks show that the MSSs in biomolecular networks are more biologically important than the MDSs in determining the system behaviours. This chapter accomplishes Objective 2 of this thesis.

Abstract

Many systems of interest in practices can be represented as complex networks. For biological systems, biomolecules do not perform their functions alone, but interact with each other to form biomolecular networks. A system is said to be controllable if it can be steered from any initial state to any other final state in finite time. Network controllability has become essential to study the dynamics of the networks and understand the importance of individual nodes in the networks. Some interesting biological phenomena have been discovered in terms of the structural controllability of biomolecular networks. Most current studies investigate the structural controllability of networks in the context of the minimum driver node sets. In this study, we analyse the network structural controllability in the context of the minimum steering node sets. We first develop a graph-theoretic algorithm to identify the minimum steering node set for a given network and then apply it to several biomolecular networks. Application results show that biomolecules identified in

the minimum steering node sets play essential roles in corresponding biological processes. The agreement of the identified steering nodes and existing research results provides a novel perspective for understanding biological systems. Furthermore, our application results indicate that the minimum steering node sets can reflect the network dynamics and node importance in controlling the networks better than the minimum driver nodes sets.

3.1 Introduction

Complex networks are ubiquitous in many scientific subjects. The last decade has witnessed an exceptional development in understanding the topology and dynamics of complex networks [112–114]. Due to the interactions among nodes in a network, perturbing some nodes can affect other nodes, which may cause the state transition of a network. Therefore, how to control networks becomes an attractive research topic.

Based on different dynamic models, several methods for controlling networks have been proposed. For nonlinear dynamic network model, Yang *et al.* [115] and Cornelius *et al.* [49] have developed different strategies to control the states of the networks by perturbing the states of some nodes. For Boolean networks, the controllability and optimal control have been investigated in [42, 116–118]. The control of linear dynamic networks attracts a lot of investigations and most of them focus on the controllability of networks. A dynamic system is completely controllable if the system can be steered from any initial state to any final state in finite time via suitable inputs. Though the controllability of linear dynamic systems is well studied by many researchers [19, 119], the criteria such as Kalman's controllability criterion and PBH criterion for determination of controllability can not be trivially applied to complex networks due to computational complexity as well as the unknown or inaccuracy of the parameters [19]. To deal with these limitations, the concept of structural controllability has been studied [20, 22, 23, 120].

Liu *et al.* [19] propose a framework to compute the minimum number of independent input control signals or driver nodes for completely structurally controlling a complex network. Every node in the minimum driver node set (MDS) is required to be actuated by an independent input control signal in order that the network is completely structurally controllable. This framework of structural controllability has inspired many recent studies in the complex network control. On the one hand, some studies extend the structural controllability framework to different linear dynamic models or different control objectives. For example, Nepusz *et al.* study the structural controllability of a switchboard dynamics (SBD) model [57]. Wu *et al.* propose the structural transittability of complex networks [56]. Other researchers investigate the output structural controllability [67, 68] and its applications in drug target identification [67].

Furthermore, several studies investigate control properties of complex networks based on Liu *et al.*'s work. The concept such as control centrality [63], control capacity [101] and control profile [107] have been proposed and studied to discover the controllability of complex networks or investigate the properties of driver nodes. In addition, the roles of nodes in MDSs of real-world networks, such as biological networks, have been

investigated [87, 100, 102]. For example, by investigating the roles of driver metabolites in the human liver metabolic network, Liu and Pan find that the driver metabolites play essential biological functions and the driver metabolites connecting different pathways are crucial in the controllability of the network. They also suggest that the environment could be important in health of human liver metabolism since the extracellular metabolites are critical driver nodes for controlling the network [102]. Liu and Pan analyse the probabilities of proteins being chosen in an MDS and compare to the roles of proteins played in a human signaling network [100]. The phenotypic properties and the genetic correlations of the neurons which act as driver nodes in neuronal network of C.elegans are investigated in [87].

In Liu *et al.*'s study [19], one driver node is corresponded to one independent input control signal and it is assumed that one input control signal can also directly actuate on other nodes outsides the MDS in the network. The node which is actuated by an input control signal is called a steering node. The minimum steering node set (MSS) consists of the minimum number of nodes which should be actuated by input control signals in order to have a network structurally controllable. Comparing the definitions between MSS and MDS, it can be seen that an MSS contains a subset as an MDS. In this study, we focus on the MSSs for completely structural controllability of complex networks.

The minimum controllability of a network refers to as the cardinality of an MSS to completely control the network. Olshevsky studies the minimum controllability problem of networks with all known values of exact parameters by formulating it as the minimum set cover problem [121]. As a result he claim that this problem is NP-hard. However, Olshevsky [122] has recently investigated the minimum structural controllability and proposed an algorithm to identifying MSS in polynomial time. In addition, Yin *et al.* [123] apply the linear integer programming method to study the minimum structural controllability of networks. In this study, we develop a novel algorithm which formulates the problem of identifying an MSS as a minimum cost maximum flow problem, which can be solved in polynomial time. Different from Olshevsky's algorithm, our developed algorithm can discover the clear relationship between MDS and MSS, which is important in some applications.

The rest of this paper is organized as follows. Section 2 presents some basic concepts of structural controllability and compares the MDSs and MSSs. Furthermore, the algorithm to identify MSSs has been introduced. Section 3 gives results of application examples of real biomolecular networks. Finally, Section 4 concludes this study and points out some directions of future work.

3.2 MSS for structural controllability

It is possible to steer a complex network from a state to another state by the application of input control signals to some nodes which are called steering nodes. If this can be done at all, there may be many different ways to do the same tasks. However, it is appealing to identify MSS required to steer a network from any initial state to any final state. In this section, one algorithm is proposed to identify an MSS in notion of structural controllability after some basic concepts and results are reviewed.

3.2.1 Network dynamic model

In this study, we consider the control of complex networks with the linear time-invariant nodal dynamic model, which can be described by the following equation:

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t) \tag{3.1}$$

where $\mathbf{x}(t) = (x_1(t), ..., x_N(t))^T$ is a state vector that describes the states of nodes in the complex network. A is an $N \times N$ state transition matrix which represents the interactions between nodes in the complex network. a_{ij} $(i \neq j)$ in the matrix A indicates the strength of the influence of node j on node i and a_{ii} is the sum of strength of self regulation and intrinsic dynamics, such as degradation, of node i. $\mathbf{u}(t) = (u_1(t), ..., u_M(t))^T$ is an input vector of M independent input control signals. The $N \times M$ matrix B is an input matrix that indicates the nodes which are directly actuated by input control signals. The dynamics of a network described by equation (3.1) is denoted as system (A, B).

3.2.2 Structural controllability conditions

According to the Kalman's controllability rank condition, system (A, B) is completely controllable if and only if the $N \times NM$ controllability matrix

$$\mathfrak{C} = \begin{bmatrix} B & AB & A^2B & ... & A^{n-1}B \end{bmatrix} \tag{3.2}$$

has the full row rank of N [34].

System (A, B) is called a structural system (A, B) when the entries in matrices (A, B) are either fixed zero or free parameters. A structural system (A, B) is completely structurally controllable if it is possible to choose the values for the free entries in matrices A and B such that the Kalman's controllability rank condition is satisfied [20].

To study the structural controllability in terms of graph theory, let G(A,B) be a digraph which contains a set of nodes $V_A \bigcup V_U$, where $V_A = \{v_1, ..., v_n\}$ and $V_U = \{u_1, ..., u_m\}$ and a set of edges $v_j \rightarrow v_i$ for $a_{ij} \neq 0$ and $u_j \rightarrow v_i$ for $b_{ij} \neq 0$. G(A) is a subgraph of G(A, B) induced by the node set V_A. The nodes in V_A correspond to the state nodes in the network and the edges between them are indicated by the state transition matrix A. The nodes in V_U represent input nodes. Each node u_i in V_U of G(A,B) corresponds to the input control signal $u_i(t)$ in $\mathbf{u}(t)$. Edges from a node u_i in V_U to nodes in V_A correspond to the ith column in the input matrix B (See Fig. 3.1).

A graph-theoretic condition for structural controllability (Theorem 3.1) has been developed in previous studies [20, 22, 23, 120]. Before introducing the Theorem 3.1, we define two following concepts which are illustrated in Fig. 3.2.

***Definition* 3.1** (Inaccessibility [20])**.** A node v_i in the digraph G(A,B) is called accessible if and only if there exists a directed path reaching v_i from the input vertices V_U, otherwise it is inaccessible.

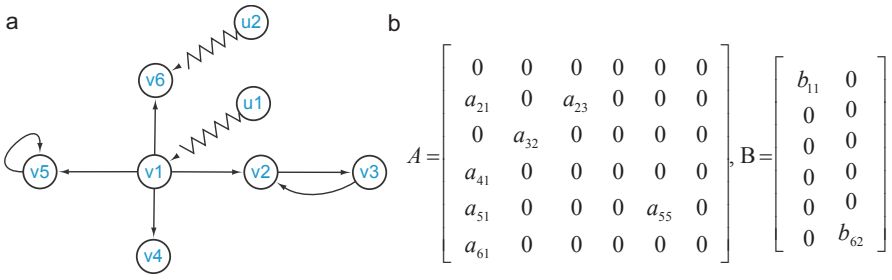


Figure 3.1: Graph representation of a system. (a): G(A,B) corresponds to system (A, B). (b): The state transition matrix and input matrix of the system (A, B).

***Definition* 3.2** (Dilation [20])**.** The digraph G(A,B) contains a dilation if and only if there is a subset S of V_A such that $|T(S)| < |S|$. Here, T(S) is the neighborhood set of S containing all nodes v_j, that there exists an oriented edge from v_j to a node in S, i.e., $T(S) = \{v_j \ | (v_j \to v_i) \in E(G),\ v_i \in S\}$. E(G) is the edge set of G(A,B). The input nodes are not allowed to belong to S but may belong to T(S). $|S|$ or $|T(S)|$ is the cardinality of set S or T(S), respectively.

***Theorem* 3.1** (Structural controllability theorem [20])**.** A structural system (A, B) is completely structurally controllable if and only if:

i) the digraph G(A,B) contains no dilation.

ii) no node in V_A is inaccessible from nodes in V_U.

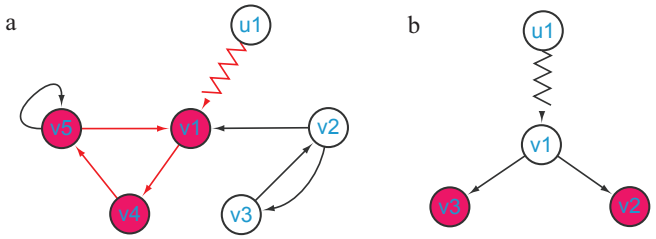


Figure 3.2: Inaccessible nodes and dilation. (a): There is no path from u_1 to v_2 and v_3, therefore nodes v_2 or v_3 are inaccessible. (b): Consider a set $S = \{v_2, v_3\}$, we have $T(S) = \{v_1\}$. Because $|T(S)| < |S|$, there exists a dilation.

According to Theorem 3.1, systems corresponding to G(A,B) in Fig. 3.2a and Fig. 3.2b are both structurally uncontrollable.

3.2.3 MDS and MSS

An MDS of a network is a minimum set of nodes to each of which should be actuated by an independent input control signal to have the network completely structurally controllable. In other words, if a system G(A,B) is structurally controllable, each node in an MDS of network G(A) should be actuated by a distinct input node in V_U. An MDS can be identified by a maximum matching-based method which has been developed by Liu *et al.* [19].

However, by only applying independent input control signals to each of nodes in an MDS, the resulting system G(A,B) can not be guaranteed completely structurally controllable. For example, in Fig. 3.3a, the red edges indicate a maximum matching in G(A), then the MDS is identified to be the nodes without incoming edges in the maximum matching, which are nodes v_1 and v_3. By applying two independent input control signals to v_1 and v_3, respectively, the resulting system is not completely structurally controllable. It is because in the G(A,B), nodes v_5 and v_6 are inaccessible from input nodes u_1 or u_2 and then the condition ii) of Theorem 3.1 is not satisfied.

Therefore, to obtain further insights of the controllability of complex networks, we proposed to investigate the MSSs. An MSS of a network is a minimum set of nodes in which each node should be actuated by an input control signal to have the network completely structurally controllable. If each of nodes in an MSS is actuated by an independent input control signal, the resulting system G(A,B) is completely structurally controllable for sure because both conditions i) and ii) of Theorem 3.1 are satisfied. In the example of Fig. 3.3b, an MSS of the network G(A) is $\{v_1, v_3, v_5\}$. By applying each of three independent input control signals to each node in the MSS, the resulting G(A,B) in Fig. 3.3b is completely structurally controllable.

However, to have a network completely structurally controllable, input control signals actuated on steering nodes are not necessarily independent. For example, an MSS of G(A) in Fig. 3.3c is $\{v_1, v_3, v_5\}$. By connecting input node u_1 to node v_1 and v_5 and connecting input node u_2 to node v_3, the resulting system is completely structurally controllable according to Theorem 3.1. In this case, nodes v_1 and v_5 are actuated by a same input control signal from u_1, and the network can be controlled by two independent input control signals.

Actually, from the definitions of MSS and MDS, we can see that each MSS contains an MDS while the MDS is a maximum subset of the MSS in which each node is connected to a distinct input node in V_U. Therefore, if G(A,B) is completely structurally controllable, each node in the MDS should be connected to an distinct input node, and the nodes in MSS while not in MDS can be connected to any input nodes. By only connecting the MDS to the input nodes, the resulting G(A,B) has no dilation but it is not guaranteed that G(A,B) has no inaccessible node. Connecting nodes in MSS to input nodes ensures that all nodes in G(A,B) are accessible.

The cardinalities of an MDS and an MSS are denoted as N_D and N_S, respectively. With the digraph representation, given a network G(A), since each node in the MDS corresponds to a distinct input node, N_D is the minimum number of input nodes which make up V_U such that the structural controllability conditions can be satisfied. N_S is the minimum number of nodes in V_A required to be connected to nodes in V_U to satisfy the structural controllability conditions. With the system described by equation 3.1, identifying an MDS is equivalent to finding a matrix B with the minimum columns such that the system (A,B) is structurally controllable. Because each column in B corresponds to an independent input control signal, the minimum number of columns of B is equals to N_D. Identifying an MSS is equivalent to finding a matrix B with not only the minimum number of columns but also the minimum number of nonzero rows such that the system

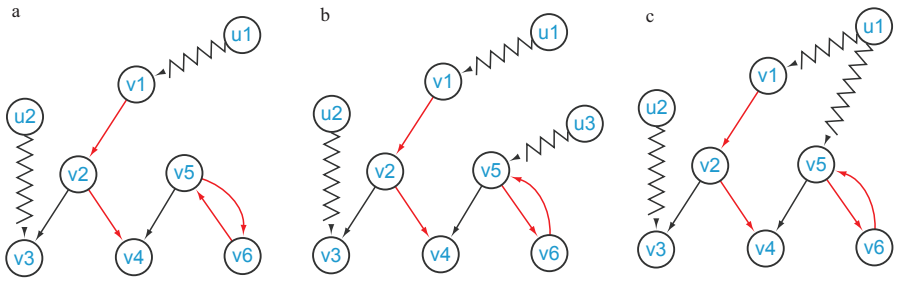


Figure 3.3: MDS and MSS. Nodes $\{v_1, v_2, v_3, v_4, v_5, v_6\}$ and the edges between them make up network G(A). The red edges indicates a maximum matching of G(A). Node sets $\{v_1, v_3\}$ and $\{v_1, v_3, v_5\}$ are an MDS and an MSS of G(A), respectively. (a): By applying independent input control signals to nodes in MDS, the resulting system G(A,B) is not completely structurally controllable. (b): By applying independent input control signals to nodes in MSS, the resulting system G(A,B) is completely structurally controllable. (c): The corresponding G(A,B) is completely structurally controllable. The steering nodes v_1 and v_5 in the MSS are actuated by a same input control signal from u_1.

(A,B) is structurally controllable. Because each nonzero row in B indicates a node in V_A which is actuated by input control signals, the minimum number of nonzero rows of B is equals to N_S.

3.2.4 Identification of MSS

Based on Theorem 3.1, Liu *et al.* map the MDS identification problem to the maximum matching problem [19]. In fact, the maximum matching can only guarantee the condition i). To identify the steering nodes, the condition ii) should be satisfied such that all the nodes can be accessible from input nodes.

We formulate the identification of MSS as a minimum cost maximum flow problem in a corresponding network G_f of a given G(A) [124] (See Fig. 3.4). To construct the network G_f, firstly we compute strongly connected components (SCCs) of G(A) [125]. If we contract each SCC to a single node, the resulting graph is a directed acyclic graph. The SCCs that corresponds to nodes without any incoming edge in resulting graph are called source SCCs. Then G_f can be constructed by the following steps:

1. Construct a bipartite graph which contains node sets $R = \{r_1, \cdots, r_n\}$ and $C = \{c_1, \cdots, c_n\}$. The nodes r_i and c_i correspond to the node i of G(A). If there is a directed edge from node i to j in G(A), there is an edge in the bipartite graph connected r_j and c_i. The capacity and cost of the edges in the bipartite graph are one and zero, respectively.

2. Add a sink node t to the bipartite graph and add edges from all nodes in C to the sink node. The capacity and cost of edges are one and zero, respectively.

3. Add a source node s to the bipartite graph.

4. Let $SCC1, SCC2, ...,$ denote different source SCCs of G(A). For each $SCCi$ that consists of more than one node, numbering its nodes as $SCCi\_1, \cdots, SCCi\_j$ $(j > 1)$:

   (a) create three auxiliary nodes a_{SCCi_1}, a_{SCCi_2} and a_{SCCi_3}

(b) add an edge from the source to a_{SCCi_1} with the capacity and cost of the edge being one and zero, respectively. Add an edge from the source to a_{SCCi_2} with the capacity and cost of the edge being $j-1$ and zero, respectively. Add an edge from a_{SCCi_1} to a_{SCCi_3} with both the capacity and cost of the edge being one. Add an edge from a_{SCCi_2} to a_{SCCi_3} with the capacity and cost of the edge being $j-1$ and zero, respectively.

(c) add edges from a_{SCCi_3} to each node of $\{r_{SCCi\_1}, \cdots, r_{SCCi\_j}\}$ with the capacity and cost of the edges being one and zero, respectively.

5. For each source $SCCi$ that consists of only one node $SCCi\_1$, add an edge from the source node to r_{SCCi_1} with both the capacity and cost of the edge being one.

6. For non-source $SCCi$ that consists of node $SCCi\_1, \cdots, SCCi\_j$, add edges from the source node to each node of $\{r_{SCCi\_1}, \cdots, r_{SCCi\_j}\}$ with the capacity and cost of the edges being one and zero, respectively.

We have the following result:

***Theorem* 3.2.** For a minimum cost maximum flow f in G_f, the MSS of the network consists of two types of nodes:

i) nodes in network whose corresponding nodes in R of G_f without the flow f passing through.

ii) if flow f passes through a_{SCCi_1}, choose an arbitrary node in $SCCi$ as a steering node.

The first type of nodes make up an MDS. To achieve structural control of G(A), each node of first type should be driven by an independent control, while the second type of nodes can be driven by any one of the input control signals for the MDS.

Before proving Theorem 3.2, we introduce the concept of power dominating set (PDS) and Lemma 3.1.

Given a network, PDS is defined to be a minimum set of nodes from which all the nodes in the network can be accessed. By assuming that each node in a network has a self link, Cowan *et al.* suggest by connecting only one input control signal to a PDS, the network is completely structural controllable [58]. In fact, by assuming the existence of self loops in G(A), there is no dilation and thus condition i) of Theorem 3.1 is always satisfied. In this special case, a matrix B with only one column can always achieve completely structural control of the networks and the nonzero entries in B are indicated by PDS.

***Lemma* 3.1.** [123] Let MDS_b and PDS_b denote a pair of MDS and PDS having the biggest intersection, then $MDS_b \cup PDS_b$ is an MSS.

Proof of Theorem 3.2: It is easy to see that a maximum flow in G_f can be mapped to a maximum matching for the identification of MDS in the work of Liu *et al.* [19]. The nodes in R which with a maximum flow passing through are one-to-one to the matched nodes in the maximum matching. Then the nodes in R without flow passing through correspond to an MDS and the cardinality of the MDS is N_D. For a given G(A), although it could has many different MDSs, N_D is the same.

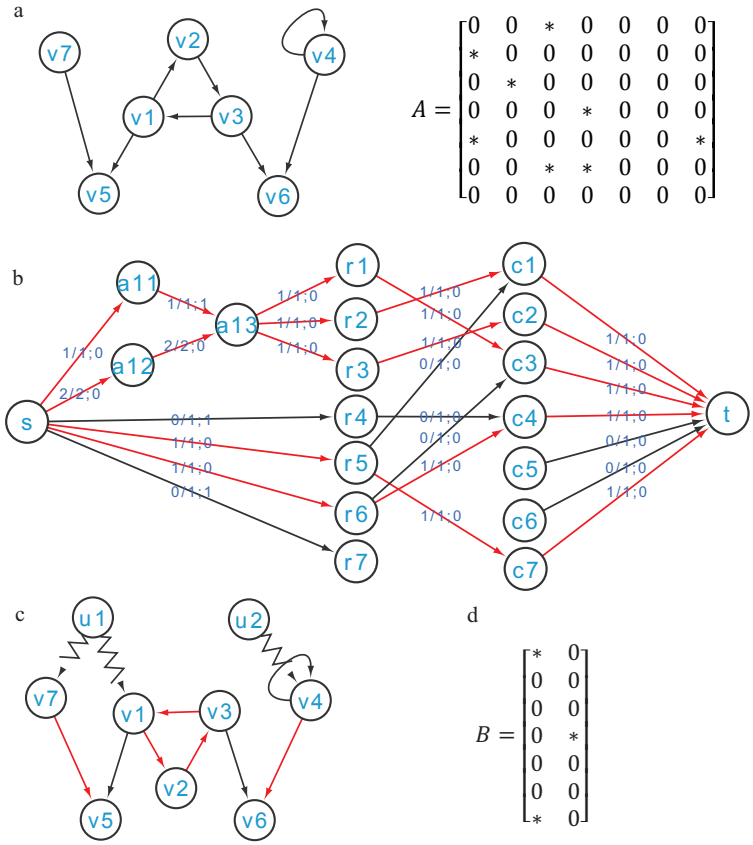


Figure 3.4: Identifying an MSS of a complex network by the minimum cost maximum flow method. (a): G(A) of a complex network and the corresponding system state transition matrix. (b): The corresponding matrix G_f and the minimum cost maximum flow. The labels represent (flow / capacity; cost) of the edges. The flow does not pass through r_4 and r_7, which suggests v_4 and v_7 make up an MDS. The flow passes through one edge with cost one, which means there is an additional steering node which should be chosen from the corresponding source SCC. (c): G(A,B) constructed based on the flow in graph (**b**) and system (A,B) is completely structurally controllable. a_{11} has flow passing through, then select any one node in corresponding source SCC as additional steering node. Here v_1 has been selected, and the MSS is identified as $\{v_1, v_4, v_7\}$. (d): The designed matrix B for having the complex network completely structurally controllable.

Considering the cost of a maximum flow in G_f, the cost c of the flow equals to the number of nodes a_{SCCi_1} which have the flow passing through, which is equal to the number of source SCCs that have no node in the MDS identified by the maximum flow. Denote PDS_{sub} as a set of nodes by arbitrarily choosing one node from each $SCCi$ whose corresponding a_{SCCi_1} have the flow passing through. Then we have $|PDS_{sub}| = c$ and $PDS_{sub} = PDS - (PDS \cap MDS)$. Therefore, it is not difficult to find out that $|MDS \cap PDS| = N_P - c$. Since the value of N_P does not change with PDS for G(A), we have $|MDS \cap PDS|_{max} = N_P - c_{min}$.

For a minimum cost maximum flow in G_f, the cost c is minimum and the corresponding MDS and PDS are a pair MDS_b and PDS_b. According to Lemma 3.1, the theorem is proved (See Fig.3.5).

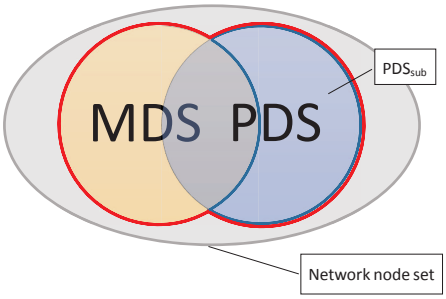


Figure 3.5: Illustration of node sets and their cardinalities. The cardinalities of MDS and PDS are N_D and N_P respectively, which are constants for given G(A). The cardinality of PDS_{sub} is c, which is the cost of a maximum flow in G_f. The cardinality of a union of any pair of MDS and PDS is equal to $N_D + c$, so the cardinality of an MSS is $N_D + c_{min}$.

3.2.5 Algorithm implementation and complexity analysis

The minimum cost maximum flow problem can be solved in $O(|f|m \log n)$ time, where $|f|$ is the amount of the flow [73]. In this study, since the G_f is specifically constructed, we can combine the Hopcroft-Karp algorithm [71] and Ford-Fulkerson algorithm [73]. The combined algorithm can solve the MSS problem with a time complexity of $O(N^{1/2}L + L * N_{sSCCb})$ in the worst case, where N, L and N_{sSCCb} represent the number of nodes, edges and the number of source SCCs which has more than one node, respectively.

According to [73], if f is a minimum-cost flow and f' is a flow by augmenting f along an augmenting path π with the minimum cost, then flow f' is a minimum-cost flow. Notice that the cost of flow on each edge is either 0 or 1, then according to the theorem in [73], finding a minimum cost maximum flow in G_f can be divided into three steps:

1. We consider a flow f_1 of G_f such that for each source SCC, there is at least one node in corresponding R which has no flow passing through. f_1 can be obtained by finding a maximum matching in a bipartite graph G_{bi}. G_{bi} is an induced subgraph of G_f which contains all the nodes in C and R'. R' is a subset of R. For each source SCC, by arbitrarily removing one node in R corresponding to the source SCC, the rest of nodes in R consist of R'. If a source SCC has only one node, remove the only node in R corresponding to this source SCC.

A maximum matching in G_{bi} is one-to-one to an eligible f_1 in G_f. Since all the edges that f_1 passes through the cost zero, f_1 is a minimum-cost flow. We denote the value of flow f_1 as v_{f1}, which equals to the number of matched nodes in set R in the maximum matching of G_{bi}. (Figure S3.6)

2. Based on the flow f_1, iteratively find an augmenting path from the source to the target with the cost of 0 and augment to the current flow. Until the iteration stops, the current flow is denoted as f_2 and the value is denoted as v_{f2}.

3. Based on the flow f_2, iteratively find an augmenting path from the source to the target and augment to the current flow. Until the iteration stops, the current flow is denoted as f_3 and the value is denoted as v_{f3}. Then f_3 is a minimum cost maximum flow, the value of the flow is v_{f3} and the cost is $v_{f3} - v_{f2}$.

Since it is known that the N_S is equal to $N_D + cost$, and we have $N_D = N - v_{f3}$ and $cost = v_{f3} - v_{f2}$, then we have N_S equals to $N - v_{f2}$. Therefore, to calculate the N_S of a network, only the first two steps need to be operated. To further identify the MDS which is a subset of MSS, the step 3 can be performed. The step 1 can be solved by Hopcroft-Karp algorithm in time $O(N^{1/2}L)$. Augmenting procedures in steps 2 and 3 can be solved by Ford-Fulkerson algorithm. Each augmenting path can be found in $O(L)$ and the iterations number will be no more than N_{sSCCb}. Generally $N_{sSCCb} \ll N$, then the time complexity is reduced to $O(N^{1/2}L)$.

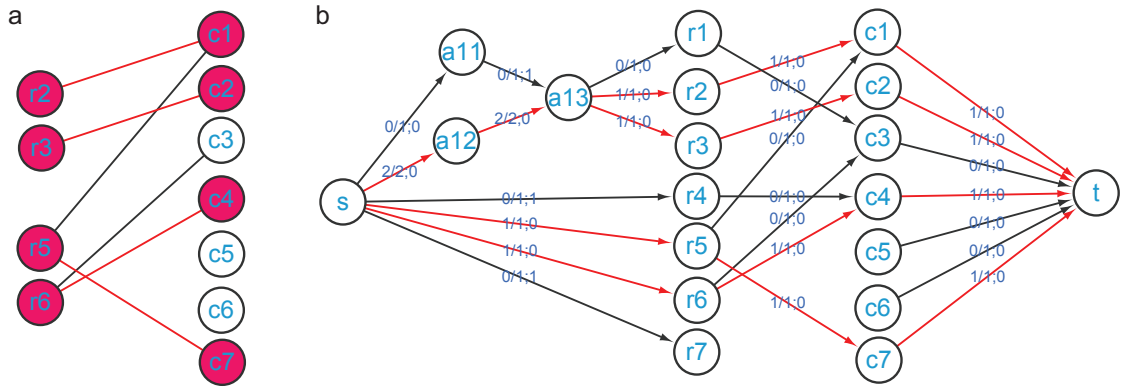


Figure 3.6: Step 1 of finding minimum cost maximum flow. (a): A corresponding G_{bi} and a maximum matching. Red edges represent the maximum matching. (b): The flow f_1 corresponds to the maximum matching in (a). Red edges represent the flow f_1.

3.3 Application results

We employ several different biological systems to demonstrate the differences between MDS and MSS. We discover that MSS is more feasible for investigating the structural controllability of real networks comparing to MDS. These examples are cell cycle networks of budding yeast [11] and fission yeast [13], Epithelial to Mesenchymal Transition (EMT) network [91] and myeloid differentiation regulatory network [37].

3.3.1 Yeast cell cycle network

The cell-cycle is a vital biological process in which the division and duplication of a cell takes place to produce its two daughter cells. For the budding yeast, Li *et al.* [11] have established a molecular network model for cell cycle of *S. cerevisiae*. For the fission yeast, Davidich *et al.* [13] have established a molecular network model for cell cycle of *S. pombe*. Applying the logic-like operations, the evolution of molecule states in different phases of cell cycles can be modeled. Sequences of molecule states can be observed which exactly matches the corresponding biological time sequences of cell cycles, from the excited G1 state (START) through S and G2 to the M phase and finally arrive at the G1 state of next cycle. For both *S. cerevisiae* and *S. pombe* models, the attractors with the largest basin size correspond to the G1 stationary states. In this study, we apply the structural controllability to analyze these two networks. For structural controllability, only regulatory relationships between molecules are required, then we study on the networks by considering both activation and inhibition edges of two original networks as simple directed edges (See Fig. 3.7).

***S. cerevisiae* network:**

The *S. cerevisiae* network consists of 11 essential molecules with 34 interactions as shown in Fig. 3.7a. We apply our method to the network and find that the minimum number of steering nodes is 1 and there is a unique MSS which consists of Cln3. This result indicates that by applying input control signal to Cln3, the network is completely controllable. In fact, the Cln3 is critical in starting the cell cycle of *S. cerevisiae*. When Cln3 is activated by external signals, signal cascades will be triggered in the network which induce the subsequent cell-cycle phases [11]. This biological observation is in agreement with the importance of Cln3 for the control of the cell-cycle network.

For MDS, by applying the method in [19] to the cell-cycle network, the minimum number of driver nodes is also 1 which could be Cln3, MBF or SBF. However, by only applying input control signal to either MBF or SBF this network cannot be completely controlled because neither of them regulate node Cln3. Therefore, one can conclude that MSS is more feasible than MDS in investigating the controllability of complex networks.

***S. pombe*:**

The *S. pombe* network consists of 9 essential molecules with 26 interactions as shown in Fig. 3.7b. Applying the maximum matching method and our algorithm, we can identify the N_D and N_S for this networks are both 2. There are 20 different sets of MDSs, however, by only applying input control signals to different MDSs, the resulting system is not always structurally controllable. Among these sets of MDSs, 6 different sets are also MSSs, which can be represented as $\{SK, v\}$, where $v \in \{Ste9, Rum1, Cdc25, Wee1/Mik1, Slp1, PP\}$. Note that all the MSSs contains node SK, because there is no directed edge from other nodes to SK, then SK should be actuated in order to have the network completely structurally controllable. According to the model of [13], the SK represents start kinases, which is a set of kinases (such as Cdc2/Cig2) that can be activated

by cell mass. SK can inhibit Ste9 and Rum1, which starts the cell-cycle process. Therefore, the fact that SK has been identified in all sets of MSSs is in agreement with the critical role that SK plays in starting a cell cycle. Besides SK, one additional node should be chosen as a steering node to form an MSS. We find that the additional node can be arbitrarily chosen in the network excepts Cdc2/Cdc13 and Cdc2/Cdc13*. In the cell cycle, concentrations of Cdc2/Cdc13 or Cdc2/Cdc13* vary during each phase transitions, which indicates the Cdc2/Cdc13 and its unphosphorylated form Cdc2/Cdc13* are essential [13]. No MSS contains these two nodes suggests that the steering nodes avoid essential nodes. Actually, these two nodes have highest degrees in the network, which suggests that steering nodes also avoid hub nodes as driver nodes discussed in [19].

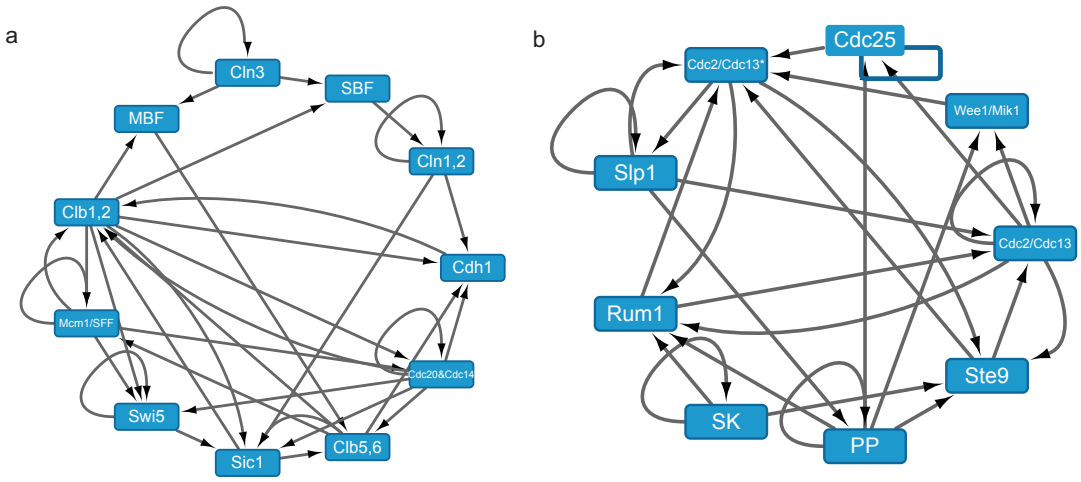


Figure 3.7: Cell-cycle network. (a): *S. cerevisiae* cell-cycle network. (b): *S. pombe* cell-cycle network.

3.3.2 EMT network

During the EMT process, cells change their genetic and trancriptomic program leading to phenotypic and functional alteration and this transition starts the metastatic dissemination [91]. To study EMT, Moes *et al.* [91] have constructed an EMT network which consists of 6 nodes and 15 interactions (Fig. 3.8). For this network, all 6 nodes are significantly differentially expressed between epithelial and mesenchymal phenotypes. Applying our algorithm, we can identify the N_S equals to 1 and node SNAI1 makes up an MSS. This result is consistent with the experimental result verified by Moes *et al.* [91] that SNAI1 can activate the transition from epithelial to mesenchymal phenotype. Besides SNAI1, we find that any node excepts CDH1 can make up an MSS of the network, too. According to the literature [91], SNAI1, MIR203 and MIR200 are key regulators of epithelial homeostasis, which also supports the ability of these nodes for controlling the network. While other two identified MSSs which consist of ZEB2 and ZEB1 respectively are deserved for the further investigation. When the maximum matching method is applied to this network, anyone of six nodes could be an MDS. However, applying input control signals to CDH1 cannot control the whole network as CDH1 does not regulate any other nodes in the network.

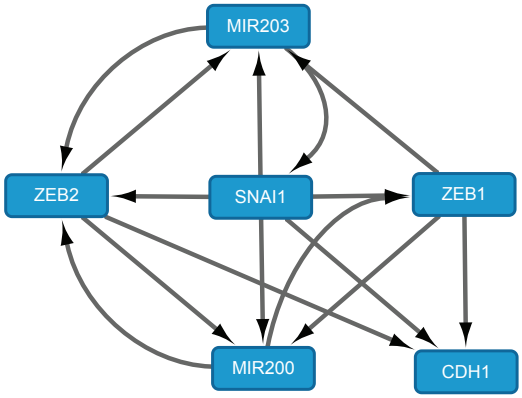


Figure 3.8: The EMT network.

3.3.3 Myeloid differentiation regulatory network

Common myeloid progenitor (CMP) is a cell state which can proliferate and differentiate into megakaryocyte-erythrocyte progenitors (MEP) and granulocyte-monocyte progenitors (GMP). The MEP and GMP could further give rise to megakaryocytes, erythrocytes, granulocytes, monocytes and others. To study these processes, Krumsiek *et al.* [37] have constructed a myeloid differentiation regulatory model network which consists of 11 transcription factors and 27 regulatory interactions (Fig 3.9). We apply our method to the network and find that the minimum number of steering nodes is 1. The MSS is unique which consists of C\EBP$\alpha$. The result suggests that by applying input control signal to C\EBP$\alpha$, the network is completely controllable. According to previous studies [37, 126], C\EBP$\alpha$ is known to be essential for development of GMP. In addition, it is known that C\EBP$\alpha$ is a key regulator of early myeloid development [127] as well as in hematopoietic stem cells [126]. The existing research results support the potential of C\EBP$\alpha$ for controlling the whole network. The minimum number of driver nodes of this network is also 1 and yet an MDS could be a set which consists of any one of C\EBP$\alpha$, GATA-1, FOG-1, PU.1, EKLF and SCL. However, by only applying input control signal to any one of GATA-1, FOG-1, PU.1, EKLF and SCL, this network cannot be completely controlled since C\EBP$\alpha$ is not regulated by other nodes. Comparing to other nodes in MDSs, C\EBP$\alpha$ plays more important role in controlling the whole network and the cell proliferation and differentiation process. Therefore, the identification of MSS is more feasible than MDS in investigating the controllability of complex networks.

3.4 Conclusion

In this work, we developed an efficient algorithm to calculate the N_S and the MSSs for completely structural controllability of complex networks. By transferring the minimum control problem into a minimum cost maximum flow problem, our approach can be interpreted meaningfully. Maximizing the flow guarantees the number of driver nodes is the minimum, while the number of steering nodes required besides the driver nodes

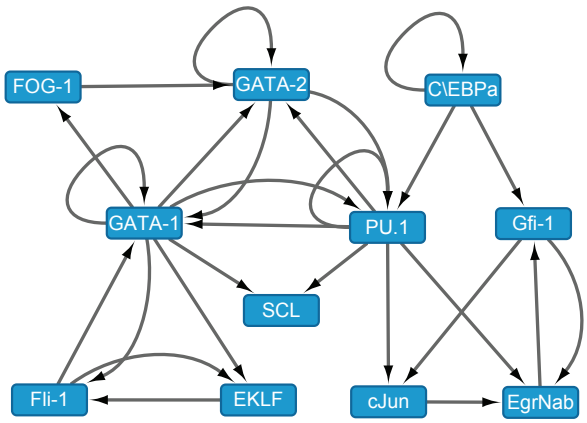


Figure 3.9: A regulatory network of myeloid differentiation.

is minimized by minimizing the maximum flow cost. Though there is a standard algorithm to solve the minimum cost maximum flow problem, our algorithm has been optimized for this particular problem.

Since only applying input control signals to MDS can not guarantee the completely structural controllability, in the realization of control strategies, identifying steering nodes is inevitable, and we always have to know which nodes should be actuated. Furthermore, we apply our method to four biological networks, the results suggest that it is more meaningful by using MSSs to investigate structural controllability of complex networks comparing to MDSs.

Taking together, our study is a supplement to the controllability of complex networks. The application results suggest that the identified steering nodes in MSSs of biomolecular networks indeed play importance roles in controlling the states of the networks. In the future, the further properties of N_S should be investigated and MSSs would be applied to biomolecular networks that control pathogenesis for drug target identification. In addition, this study only focuses on the controllability of networks while the inputs u and control time are not taken into consideration. Therefore, input control signal design for optimally controlling network could be one direction of our future work for practical applications.

Acknowledgment

This research is supported by Natural Sciences and Engineering Research Council of Canada (NSERC), the National Natural Science Foundation of China under Grant No.61428209, and Chinese Scholarship Council (CSC).

This chapter is reproduced by permission of the Institution of Engineering & Technology.

4 Network output controllability-based method for drug target identification

Published as: L. Wu, Y. Shen, M. Li, and F.-X. Wu, "Network output controllability-based method for drug target identification," *IEEE Transactions on NanoBioscience*, vol. 14, no. 2, pp. 184-191, 2015.

In the previous chapter, an algorithm was proposed to identify the MSS for complete controllability of a biomolecular network. Nodes in the MSS have important biological functions. However, the MSS only reflects complete controllability, which is the controllability of the whole network. In many situations, controlling a portion of nodes instead of the whole network is more practical.

In this chapter, controllability of a subset of nodes is formulated as the problem of network output controllability. An algorithm is developed to identify steering nodes for output controllability of a network. The method can be used to identify drug targets based on biomolecular networks: disease related biomolecules correspond to a subset of nodes which can be controlled while the identified steering nodes are potential drug targets. This chapter fulfills Objective 3 of this thesis.

Abstract

Biomolecules do not perform their functions alone, but interactively with one another to form biomolecular networks. It is well known that a complex disease stems from the malfunctions of corresponding biomolecular networks. Therefore, one important task is to identify drug targets from biomolecular networks. In this study, drug target identification is formulated as a problem of finding steering nodes in biomolecular networks while the concept of network output controllability is applied to the problem of drug target identification. By applying control signals to these steering nodes, the biomolecular networks are expected to be transited from one state to another. A graph-theoretic algorithm is proposed to find a minimum set of steering nodes in biomolecular networks which can be a potential set of drug targets. Application results of the method to real biomolecular networks show that identified potential drug targets are in agreement with existing research results. This indicates that the method can generate testable predictions and provide insights into experimental design of drug discovery.

4.1 Introduction

The last decade has witnessed the exceptional development in biomolecular interaction data and most attention has been paid to biomolecular networks [128]. Cellular systems can be represented as biomolecular networks which are graphs consisting of nodes and edges. According to various levels of interactions between biomolecules, there are different types of biomolecular networks in cellular systems. For example, gene regulatory networks (GRNs), metabolic networks and protein and protein interaction (PPI) networks represent the directed or undirected interactions between biomolecules [129].

Network science is an interdisciplinary academic field that investigates the topology and dynamics of complex networks [102, 130]. The network studies have become powerful tools in the field of biology to discover the properties and understand the functions of biological systems. For example, network centralities are quantitative indices for assessing the position of a node relative to the other nodes, and have been used to elucidate the functional roles of components in different biomolecular networks [129]. Studies have revealed that the centralities of biomolecules correlate with their essentialities and many researchers have used centralities to distinguish essential components (i.e. essential proteins) or reactions in biomolecular networks [131–135]. In addition, many other network-based methods can be applied to understand biological systems, such as network clustering, which has been used to detect functional modules in PPI networks [9, 136, 137], and network alignment, which has been used to query a subnetwork or pathway that was previously known to be functional modules from a given large network or database [138].

For the problem of drug target identification, the single-target approach in drug discovery was dominant for a long time [139]. However, there are many limitations for drug design against single-target in the aspects of drug efficiency and safety. For the aspect of efficiency, drugs in clinical treatment may not be as efficient as predicted in the experiment because of the interactions between pathways in biomolecular networks. Many biomolecular networks are robust so that the change of a single target would be offset by the interactions in the networks, which makes the phenotypes of the biological systems unchanged [140]. On the other hand, side effects often come from the undesired effects of drugs on the biomolecular networks. A single-target drug may affect the states of non-target biomolecules and cause unexpected effects that can not be eliminated. To overcome the weaknesses of single-target drugs, multi-target drug design has attracted growing attention in recent years. Systems biology, which uses network-based approaches to study biomolecular networks as whole systems, plays a vital role in drug design [128, 141, 142].

Past studies have made great progresses in discovering disease-related information based on network approaches, such as identification of disease biomolecules and drug targets. Several network-based methods have been developed to identify disease biomolecules, including linkage-based, module-based, diffusion-based, and Markov random field-based methods [109, 142–146]. These approaches are based on different assumptions on the properties that disease biomolecules may have in biomolecular networks. For example, the linkage-based method tends to associate a protein with a disease if its directly interacted proteins relate to the

disease [143]. By this method, Janus kinase 3 (JAK3) has been correctly predicted to involve in combined immunodeficiency syndrome.

Drug targets are combinations of biomolecules in the networks and by changing their states biomolecular networks in abnormal states can be driven to healthy ones. To identify drug targets, some approaches which focus on the topology of biomolecular networks have been proposed. Csermely *et al.* [140] introduce a network-based method for identifying multi-target drug identification. Hwang *et al.* [147] propose a concept named bridging centrality for each node and suggest that the biomolecules with high bridging centralities are potential drug targets. Besides the studies which only focused on the static topological properties of biomolecular networks, there are other approaches based on the dynamics of biomolecular networks. These approaches are based on different dynamic models of biomolecular networks and intend to change the states of biomolecular networks by directly affecting the states of identified drug targets. In the dynamic-based approaches, the state changes of disease or non-disease biomolecules are considered as drug efficiency or toxicity. Yang *et al.* [115] construct an arachidonic acid metabolic network (AAnetwork) in which all the interactions between biomolecules are expressed by ordinary differential equations. Based on the model, not only have optimal target sets been identified, but also mechanisms for the side effects of existing drugs NSAIDs and Vioxx have been found. Li *et al.* [15, 16] use a flux balance analysis based linear program model to discover drug targets, which does not require much detailed knowledge of network dynamics compared to Yang's method. Wu *et al.* [148] develop a network dynamic model to identify effective drug combinations and successfully detect the combination of Metformin and Rosiglitazone, which is actually Avandamet, an effective drug to cure Type 2 diabetes.

The control of networks is a central issue in network science [19, 102]. Several researchers have applied network control strategies to the identification of drug target sets in biomolecular networks. In the view of control theory, drug targets in biomolecular networks can be interpreted as steering nodes. By applying signals to the set of steering nodes the networks are expected to be steered to desired states. Based on the Boolean networks model, Kim *et al.* [42] apply the genetic algorithm to search for a minimum steering node set. Based on nonlinear dynamic network model, Cornelius *et al.* [49] study the T-cell survival signalling network that governs the development of T-LGL leukemia and rank the nodes according to their potential to be steering nodes.

In this study, we apply the concept of structural controllability [19, 20, 22, 24] to the problem of drug target identification. The definition was firstly proposed by Lin in 1974 and several researchers generalized Lin's result or proved it using different methods [22, 23, 120]. The structural controllability only concerns whether there are interactions between nodes in a network and does not consider the strength of the interactions. Liu *et al.* apply the theory of structural controllability to different types of real networks and determine the minimum set of so-called driver nodes, to which applied by input signals, the networks can be driven to any desired state in finite time [19]. Wu *et al.* [56] study the structural transittability of complex networks. Unlike the complete controllability which concerns the ability of a system that transits from any state to any

other state, transittability concerns the ability of the system that transits between two specific states. To steer a system from an abnormal state to a healthy state, using the transittability needs less steering nodes and has smaller state space affected than complete controllability. Instead of focusing on nodal dynamic model, Nepusz *et al.* [57] propose an edge dynamic model, which has been called the switchboard dynamics (SBD), and the controllability of networks with SBD has been studied. In addition, based on the structural controllability, Liu and Pan investigate the roles of driver metabolites in the human liver metabolic network and suggest such metabolites play essential biological functions [102].

In this study, we present a new drug target identification method based on the structural output controllability of complex networks. Section II presents the formulation of drug target identification and the algorithm to find out the minimum set of steering nodes which are potential drug targets. Section III gives results of verification experiments on real biomolecular networks. Finally, section IV concludes this study and points out some directions of future work.

4.2 Problem Formulation

If a system is completely controllable, it can be steered from any state to any other state via appropriate inputs. For a biological system at an abnormal state, if perturbing some biomolecules can affect other biomolecules and steer the system to a healthy state, these perturbed biomolecules can be considered as drug targets. Thus the problem of identifying drug targets can be formulated as the problem of finding sets of steering nodes in systems. By applying the control signals to these nodes, the systems can be steered from undesired states to other desired states.

4.2.1 Dynamic model of biomolecular networks

Even though complex dynamic processes are nonlinear, the controllability of nonlinear systems is structurally similar to that of linear systems in many aspects [19, 149]. Actually, to ultimately develop the control strategies for complex nonlinear networks, a necessary and fundamental step is to investigate the controllability (especially structural controllability) of complex networks with linear dynamics [56]. The drug target identification based on the structural controllability can provide a preliminary understanding of potential drug targets. In this study, we use the linear time-invariant nodal dynamic model, which is the most popular and canonical model, to represent the dynamics of a biomolecular network with n nodes:

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t) \tag{4.1}$$

where $\mathbf{x}(t) = (x_1(t), ..., x_n(t))^T$ is a state vector that describes the states of nodes in the biomolecular network. From the perspective of a biomolecular network, states of nodes can be concentrations of metabolites or enzymes in a metabolic network or can be expression levels of genes in a gene regulatory network. A is an $n \times n$ state transition matrix which represents the interactions between nodes (biomolecules) in the

biomolecular network. Each entry a_{ij} in the matrix A indicates the strength of the influence of biomolecule j on biomolecule i. The $n \times m$ matrix B is an input matrix that corresponds to the steering nodes and b_{ij} indicates the strength of an external input j to biomolecule i in the network. $\mathbf{u}(t) = (u_1(t), ..., u_m(t))^T$ is an input vector of m external inputs and each external input corresponds to a column in B. External inputs to the biomolecular network can be considered as different types of stimuli to the network, such as environment effects and drugs. In this study, we only consider drugs as external inputs which are control signals.

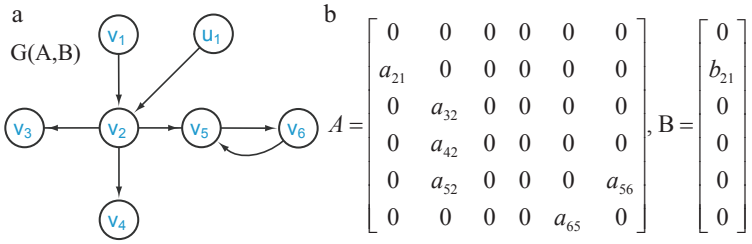


Figure 4.1: (a): A graphic representation of a system. (b): The corresponding matrices (A, B) of the system.

We use matrix pair (A, B) to represent the system described by equation (4.1). A system with matrices (A, B) can be represented as digraph $G(A, B)$. $G(A, B)$ is a digraph which contains sets of nodes $V_A \bigcup V_U$, where $V_A = \{v_1, ..., v_n\}$ and $V_U = \{u_1, ..., u_m\}$. The edge set is the pairs $v_j \to v_i$ for $a_{ij} \neq 0$ and $u_j \to v_i$ for $b_{ij} \neq 0$. The nodes in V_A and the edges between them make up the biomolecular network. The nodes in V_U represent input nodes. Edges from a node u_i in V_U to nodes in V_A correspond to a column in the input matrix B. Fig 4.1 shows an example of the graphic representation.

4.2.2 Completely structural controllability and structural output controllability

Suppose the control signals $\mathbf{u}$ can be chosen arbitrarily, input matrix B should be well designed in order to steer a biomolecular network from an abnormal state to a healthy state. Finding drug targets becomes the problem of finding an input matrix B. To measure the ability of steering the networks to healthy states, the concept of controllability has been used.

In control theory, a system is completely controllable if for any initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ and any other final state $\mathbf{x}_1$, there exists a finite time t_f and inputs $\mathbf{u}(t)$, such that $\mathbf{x}(t_f) = \mathbf{x}_1$ [149, 150]. According to the Kalman's controllability rank condition, system (4.1) is completely controllable if and only if the controllability matrix

$$\mathfrak{C} = \begin{bmatrix} B & AB & A^2B & ... & A^{n-1}B \end{bmatrix} \tag{4.2}$$

has full row rank n [34]. To completely control a system, the matrix B should be properly designed to satisfy the Kalman's condition, which is equivalent to find an appropriate set of steering nodes in the network.

However, there are too many nodes that control signals should be applied to in order to make a biomolecular network completely controllable. Specifically, in gene regulatory networks, independent control signals should directly apply to at least about 80% of nodes to completely control the networks [19]. For the drug

target identification problem, it is difficult and unnecessary to completely control the networks. In reality, we more concern about the states of the nodes that will relate to diseases or side effects. So we take the states of disease biomolecules and side effects causing biomolecules into consideration. The output controllability of networks can measure the controllability of a set of nodes in the networks, which motivates us to study the output controllability of biomolecular networks in this study.

The outputs of a network (A, B) can be expressed by the following equation:

$$\mathbf{y}(t) = C\mathbf{x}(t) \tag{4.3}$$

where $\mathbf{y}(t) = (y_1(t), ..., y_p(t))^T$ is an output vector of the network and each entry represents an output. The outputs of the network are linear combinations of node states in the networks represented by the $p \times n$ matrix C. In this research, we assume that for the output matrix C, there is one and only one entry which has nonzero value in each row of C while other entries are zero, and the nonzero entries are in different columns. With this assumption, each entry in $\mathbf{y}(t)$ represents the state of one node that is indicated by a corresponding row of C. The outputs $\mathbf{y}(t)$ are the states of a set of biomolecules in the biomolecular network. A system described by equations (4.1) and (4.3) is denoted by matrix triplet (A, B, C).

A system is output controllable if for any initial output vector $\mathbf{y}_0 = \mathbf{y}(t_0)$ and any other final output vector $\mathbf{y}_1$, there exists a finite time t_f and inputs $\mathbf{u}(t)$, such that $\mathbf{y}(t_f) = \mathbf{y}_1$. For a dynamic system model described by equations (4.1) and (4.3), the $p \times mn$ output controllability matrix is defined as:

$$\mathfrak{o}\mathfrak{C} = \begin{bmatrix} CB & CAB & CA^2B & ... & CA^{n-1}B \end{bmatrix}. \tag{4.4}$$

Based on the control theory, system (A, B, C) is output controllable if and only if rank $(\mathfrak{o}\mathfrak{C}) = p$ [69].

In the problem of drug target identification, we concern about the states of the disease biomolecules as well as the biomolecules whose state changes would lead to side effects. The objective of drug design is to make the states of disease biomolecules healthy while keeping the state changes of side effect-causing biomolecules minimal. So we consider the outputs of the biomolecular networks are the states of these two types of biomolecules, which determine the output matrix C. For a biomolecular network represented by matrix A and the output matrix C has been identified by the specific nodes, the problem of drug target identification can be formulated as finding a matrix B that makes the system (A, B, C) output controllable. Fig. 4.2 shows the framework of the drug target identification process.

Though the rank of matrix $\mathfrak{C}$ ($\mathfrak{o}\mathfrak{C}$) can judge whether a system is completely (output) controllable, it is difficult to calculate the rank directly. When n becomes large, the calculation of the rank of $\mathfrak{C}$ would be time consuming and hardly be accurate. In addition, for most biomolecular networks, we often only know whether there is an interaction between two biomolecules, but don't know the strength of the interaction between them. So we can only judge whether an entry in matrix A is zero or not. A matrix is said to be a structural matrix if its entries are either fixed zeros or independent free parameters. System (A, B, C) is called a structural system if A, B and C are structural matrices. "Structural controllability" is a concept which measures the controllability of structural systems [20].

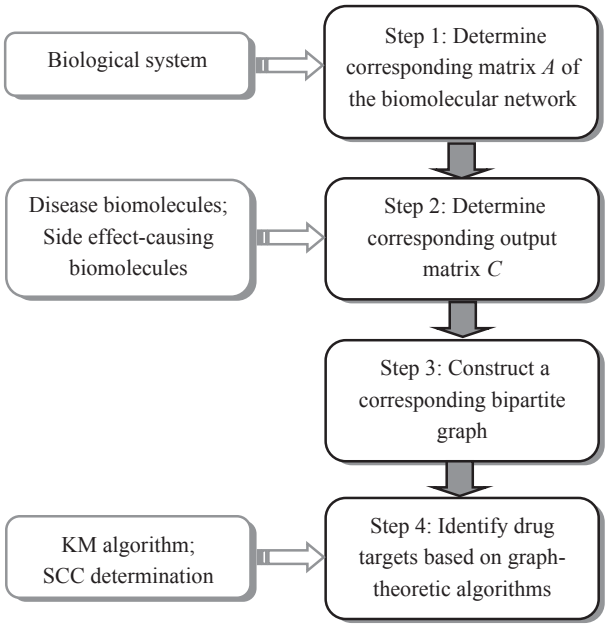


Figure 4.2: Framework of the method. The right blocks illustrate the processes of drug target identification and left blocks indicate the knowledge needed.

For a structural system (A, B), the dimension (rank of controllability matrix $\mathfrak{C}$) of its controllable subspace varies as a function of free parameters in matrices A and B. By arbitrarily choosing the value of free parameters in A and B, the dimension of controllable space of structural system (A, B) can reach a maximum value. We define the maximum value as the *generic dimension of the controllable subspace* of structural system (A, B) and denote it by $GDCS(A, B)$ [56]. Similarly, the maximum value of rank of $\mathfrak{oC}$ achieved by arbitrarily choosing the free parameters is defined as the *generic dimension of the controllable output subspace* of structural system (A, B, C) and denoted by $GDCOS(A, B, C)$. A structural system (A, B, C) is called completely structurally controllable if $GDCS(A, B) = n$ and structurally output controllable if $GDCOS(A, B, C) = p$.

To analyze the completely structural controllability, a graph-theoretic method has been proposed to calculate the $GDCS(A, B)$ [22]. In a graph, a sequence of edges $\{(v_1 \to v_2), (v_2 \to v_3), ..., (v_{k-1} \to v_k)\}$ is called a simple path if all the nodes $\{v_1, v_2, ..., v_k\}$ are distinct. If $v_1 = v_k$ and other nodes are distinct, the path is called a simple cycle. A node v_i in V_A is called inaccessible from input nodes if and only if there are no paths that can reach it from nodes in V_U. To illustrate the graph-theoretic method, we define a *path and cycle covering* to be a set of simple paths starting from input nodes in V_U and simple cycles that are accessible from input nodes where all the paths and cycles have no nodes in common.

In digraph $G(A, B)$ of a structural system (A, B), the value of $GDCS(A, B)$ is equal to the maximum number of nodes in V_A that can be covered by a *path and cycle covering* [20, 22]. Based on this result, Liu *et al.* use a maximum matching algorithm to find out a set of V_U that has the minimum number of nodes, which is equivalent to, construct a structural matrix B that has the minimum number of columns, to make

the system (A, B) completely structurally controllable [19].

However, the $GDCOS(A, B, C)$ can not be calculated accurately by any graph-theoretic method yet. Murota and Poljak [24] have developed a method to calculate the upper and lower bounds of $GDCOS(A, B, C)$. In this study, to make sure the designed matrix B can make the system structurally output controllable, we use the lower bound. Based on our assumption, each row in C corresponds to one node in the biomolecular network which is a node in V_A. p rows of C correspond to p different nodes which is a subset of V_A and denoted as V_O. Notice that if every node state is considered as an output, the output controllability reduces to the complete controllability. According to [24], the lower bound of $GDCOS(A, B, C)$ is the maximum number of nodes in V_O that can be covered by a *path and cycle covering* in $G(A, B)$. Similar to Liu *et al.* [19] whose work is to find a matrix B that makes the system completely structurally controllable, we study the structural output controllability of complex networks and apply it to identify drug targets from biomolecular networks. We develop a new algorithm to identify a minimum steering node set (drug targets) needed to control the outputs of the system based on the calculation of the lower bound of $GDCOS(A, B, C)$.

4.2.3 Method description

Given a biomolecular network whose structural matrices A and C have been determined, we would like to develop an algorithm to find a matrix B which makes the $GDCOS(A, B, C)$ equal to p. Since we expect less biomolecules to be drug targets for certain disease, which makes more feasible for drug production, the algorithm should be designed to minimize the number of steering nodes. In addition, the $GDCS(A, B)$ measures the dimension of controllable subspace of the structural system. The smaller value of $GDCS(A, B)$, the lower dimension of state space will be affected, and thus the less chance there are side effects [56]. Then the matrix B identified by the algorithm should ensure the structural output controllability of the system and try to make $GDCS(A, B)$ small. Since there is no algorithm to calculate the $GDCOS(A, B, C)$ accurately, the developed algorithm makes sure the lower bound of $GDCOS(A, B, C)$ equals to p, which guarantees that the system is structurally output controllable. In this study, we assume that each column in B contains one and only one nonzero entry, which means if we choose a node as a drug target, the node can be controlled independently. In graph-theoretic representation, every node $u_i \in V_U, i = 1, .., m$ can connect with one and only one node $v_j \in V_A, j = 1, ..., n$.

Firstly we construct a weighted bipartite graph that corresponds one on one to the network A and the output C. Nodes $r_i \in V_R$ and $c_j \in V_C$ in the bipartite graph correspond to row i and column j of matrix A, respectively. $S \subseteq V_A$ denotes the set of nodes within which the drug targets will be selected and $V_S \subseteq V_R$ denotes corresponding nodes of S in V_R. For example, we can select a set of nodes which represent enzymes in a metabolic network as set S. There are k candidate input nodes $u'_i \in V_{U'}$ and each u'_i connects to a distinct node in V_S, where k is the number of nodes in V_S and $V_{U'}$ denotes the candidate input node set in the bipartite graph.

The edges and the weights of the bipartite graph are defined as follows:

$$w_{r_i c_j} = \begin{cases} 1 & a_{ij} \neq 0 \text{ and } v_i \in V_O \\ 0 & (a_{ij} \neq 0 \text{ or } a_{ij} = 0,\ i = j) \text{ and } v_i \notin V_O \end{cases} \tag{4.5}$$

and for every $r_{s_i} \in V_S$, where s_i indicates that the ith node in V_S is the $s_i th$ node in V_R, there are edges

$$w_{r_{s_i} u'_i} = \begin{cases} \varepsilon \times rcc_{s_i} & v_{s_i} \in V_O \\ \varepsilon \times rcc_{s_i} - 1 & v_{s_i} \notin V_O \end{cases} \tag{4.6}$$

where $rcc_{s_i} \in \{1, 2, ..., n\}$ is the ranking of control centrality of node v_{s_i} in $G(A, B)$ [63], the node with largest control centrality has $rcc_{s_i} = 1$. ε is an arbitrarily small positive number and less than $2/k(k+1)$ for a network with k nodes in S. Therefore we can ensure that for the score any maximum weight matching, the sum of terms that containing ε is less than 1. For simplicity, ε can take the value of 0.001, 0.0001, 0.00001 or the like.

Then we use Kuhn-Munkres (KM) algorithm to find out the maximum weight complete matching of the constructed bipartite graph [151]. The maximum weight complete matching means every node in V_R has been matched to a distinct node in V_C or V'_U while the sum of the weights of edges in the matching is maximum. Each node in V_R that matches a node in V'_U is considered as a steering node in the corresponding digraph and can be regarded as a potential drug target.

If there is no complete matching for the constructed bipartite graph, the KM algorithm will not work. This means system is not able to be output controllable with the steering nodes chosen from set S and the drug target combination can not be found in the S by our method.

A complete matching of the bipartite graph corresponds to a covering consisting of simple paths starting from input nodes and cycles that covers all the nodes in V_O. If r_i matches c_i and $w_{r_i c_i}$ is zero, node v_i would be a node uncovered by the covering in $G(A, B)$. If r_i matches $c_j (i \neq j)$ or $i = j$ while $w_{r_i c_i} = 1$ in the complete matching, there is an edge from v_j to v_i in the covering. For $v_i \in V_O$, if $a_{ii} = 0$, there is no edge between r_i and c_i in the corresponding bipartite graph, then r_i can not match c_i and v_i must be covered by the corresponding covering. If r_{s_i} matches a node in V'_U, v_{s_i} is a steering node which is directly connected to an input node in the corresponding $G(A, B)$.

It can be proved that the number of nodes in V_R matched in V'_U is minimum (see Section 4.5), which minimizes the number of steering nodes. Control centrality measures the controllable subspace of a node v_i as steering node [63]. It is equal to the $GDCS(A, b)$, where b is a column vector and the only non-zero entry is the ith one. To reduce the $GDCS(A, B)$, the node v_{s_i} with small control centrality has large rcc_{s_i} and therefore has high priority to be matched in a maximum weight matching. Though priority of nodes to be steering nodes does not ensure the $GDCS(A, B)$ minimum, a node with smaller control centrality is more likely to be considered as a steering node, which reduces the $GDCS(A, B)$ in many cases.

After the maximum weight matching, we search for all the nodes in V_O covered by cycles and then check whether if all these cycles are accessible from input nodes. If so, the covering is a *path and cycle*

covering. Otherwise, we can add more input nodes to make the inaccessible cycles which contain nodes in V_O accessible from input nodes. This can be accomplished by a graph searching algorithm based on strong connected components (SCCs) identification. We choose some nodes from higher hierarchy SCCs as drug targets to make all cycles which contain nodes in V_O accessible from input nodes [58, 63], which guarantees that the system is structurally output controllable.

Fig. 4.3 is an illustrative application of the proposed method to the biomolecular network.

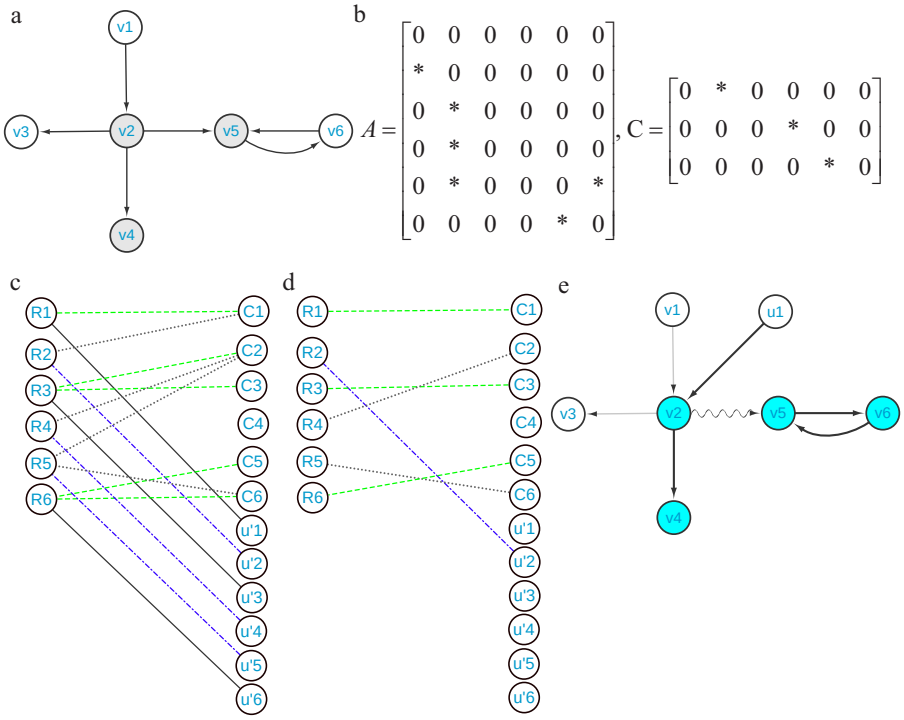


Figure 4.3: An illustrative example. (a): A biomolecular network. The outputs of the system are the states of tinted nodes. (b): Step 1, construct a corresponding structural matrix A of biomolecular network in Fig. 4.1. By assuming that v_2 and v_5 are related to a certain disease and the abnormality of v_4 will cause serious side effects, construct the corresponding structural matrix C as step 2. "$*$" in matrices A and C represents the free parameters while "0" represents the fixed parameters. (c): Step 3, the weighted bipartite graph. The weights of dot lines, dash lines, dash dot lines and solid lines are 1, 0, $\varepsilon \times rcc_{s_i}$ and $\varepsilon \times rcc_{s_i} - 1$, respectively. (d): Step 4, one of the maximum matching results. (e): Simple path and cycle in original graph corresponds to d). Because the cycle $v_5 \to v_6 \to v_5$ is accessible from drug target node v_2, no more drug targets node needed. The identified drug target is node v_2.

4.3 Results

We employ several biomolecular networks with different phenotypes or disease states in order to demonstrate the applicability and feasibility of our method.

4.3.1 Results from drug discovery-relevant human networks

AAnetwork: The AAnetwork (arachidonic acid metabolic network) is an inflammation-related work model constructed by Yang *et al.* [115]. Inflammation is a protective attempt by the organism to harmful stimuli, such as pathogens, damaged cells, or irritants. Inflammation mediators are chemical substances which cause or participate in inflammation. Some metabolites such as COX-2 and 5-LOX which play key roles in generating inflammation mediators are considered as typical drug targets. However, single-target anti-inflammation drugs always have side effects on the treatment. Attempting to decrease the drug toxicity, Yang *et al.* study the multiple target optimal interventions on the AAnetwork. In their study, the dynamics of AAnetwork is represented by a group of ODE functions based on enzymatic kinetic and the intervention solutions are obtained by Monte Carlo simulated annealing stimulation.

In this study, we construct a network with 38 nodes and 69 edges (see Fig 4.4) based on the ODE functions in the AAnetwork model. Enzymes and metabolites are considered as nodes in the network and there are three types of edges: from reactants to products, from enzymes to products and from metabolites to enzymes. The edges from metabolites to enzymes represent the inhibition, activation and upregulation to the enzymes. In the AAnetwork, the concentrations of LTB4 and PGE2 are related to the inflammation and PGI2 and TXA2 are related to cardiovascular and bleeding side effects. Then states of these four corresponding nodes are considered as output nodes in V_O of the network when applying our proposed method, which takes both efficiency and safety into consideration. Similar to stage two in [115], that drug target combinations are selected from nodes PLA2, LTA4H, 5-LOX, PGES, COX-2 and COX-1, we consider these nodes in the set S.

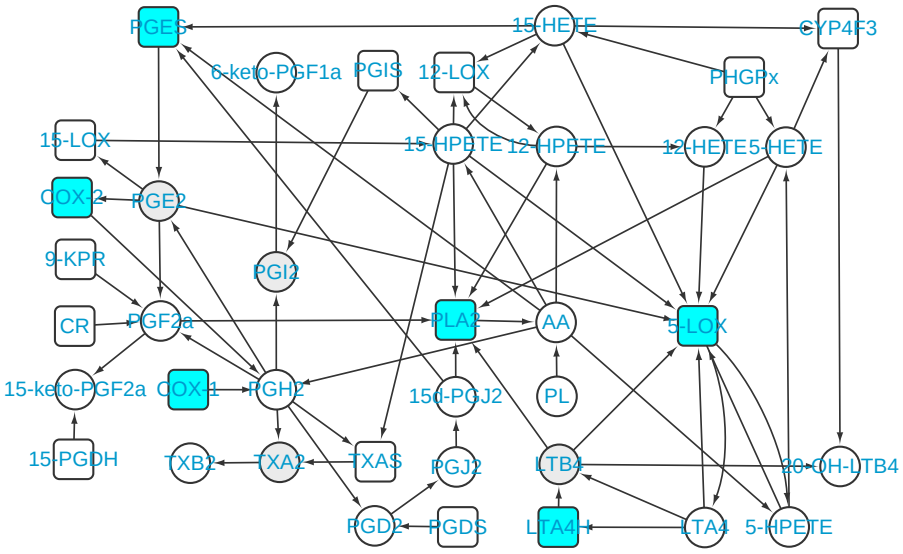


Figure 4.4: AAnetwork. The nodes represented by circles correspond to metabolites and the nodes represented by round rectangles correspond to enzymes in the AAnetwork. The tinted round rectangle nodes make up the set V_O which represents the outputs of the network. The tinted circle nodes make up the set S.

The proposed algorithm finds that a steering node set consisting of 5-LOX and PGES make the network output controllable and the corresponding $GDSC$ is 20, which is in agreement with the knowledge that 5-LOX is a standard drug target. The identified potential drug target combination by our algorithm is not the unique set that satisfies the output controllability. Therefore we verify whether the 23 sets of drug target combinations identified by Yang *et al.* can make the network system output controllable. We find that 21 sets of their results satisfied the output controllability requirement, which supports that our proposed method can give a preliminary view of the potential drug targets.

In addition, drugs such as Aspirin, Celecoxib and Vioxx which target at COX-1 and COX-2 can not make the AAnetwork output controllable. In fact, these drugs are considered to be lack of efficiency or will cause cardiovascular or bleeding problems. The drug Licofelone, which target at 5-LOX, COX-1 and COX-2 can make the AAnetwork output controllable, is an effective anti-inflammation drug without causing cardiovascular or bleeding problems. The $GDCS$ of selecting 5-LOX, COX-1 and COX-2 as drug targets is 25, which is larger than the $GDCS$ of selecting 5-LOX and PGES. In fact, PGES converts PGH2 to PGE2 which is functionally coupled with COX-2. PGE2 is also important for the formation of cytosolic enzyme which is functionally coupled with COX-1 and COX-2 [152]. So the drug targets of 5-LOX and PGES identified by the proposed method can provide a reasonable suggestion for the future drug development research.

4.3.2 Results from H.sapiens pathways from KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource that deal with genomes, biological pathways, diseases, drugs, and chemical substances [153]. When applying our method to these metabolic pathways, we consider both metabolites and enzymes as nodes in the constructed networks and select drug targets from the nodes representing enzymes.

hsa00340: Severity of Parkinson disease is affected by the level of pros-methylimidazoleacetic acid in the histidine metabolism network (hsa00340). Rasagiline mesylate (D02562), which targets on monoamine oxidase, is used as an anti-parkinsonian drug. Two studies [15, 154] are on drug target identification and the biological significance of their methods has been verified by the experiments on this histidine metabolism network. To verify our proposed method, we choose methylimidazoleacetic acid and methylimidazole acetaldehyde as the output of the system, which is the same as the experiments in [15, 154]. We have found that the optimal drug target is monoamine oxidase which is the identical to the results in [15, 154]. If a structural system is structural controllable, almost all numerical systems which have the same structure as the structural system are controllable [20, 24]. This fact can explain our result by only using network structure is the same as those in [15, 154].

hsa00230: Febuxostat (D01206) is a drug used to treat hyperuricemia by inhibiting xanthine dehydrogenase (EC:1.17.1.4) and xanthine oxidase (EC:1.17.3.2). Excess of urate will lead to hyperuricemia. We have applied our algorithm to the related metabolic pathway purine metabolism network (hsa00230) and defined the state of urate as the system output. We find that enzyme EC:2.4.2.16 is the optimal drug target based

on our model, which is different from the existing drug of febuxostat and the result from Li *et al.* [15]. Li *et al.* suggest that the drug target set consists of three enzymes which are EC:1.17.1.4, EC:1.17.3.2 and EC:2.4.2.16 and urate can be affected by all these enzymes. However, according to our model, we consider that by controlling only one of these three enzymes we can control the state of urate. Choosing EC:2.4.2.16 can achieve the smallest $GDCS$, which means its side effect is minimum. In fact, choosing EC:1.17.1.4 and EC:1.17.3.2 as targets will not only lead to a large controllable subspace, but also affect many other unrelated compounds in the network comparing to choosing EC:2.4.2.16, which means much more side effects.

4.4 Conclusion

Identifying drug targets is a major challenge in new drug development where both drug efficiency and safety should be taken into consideration. Previous network based approaches have made some achievements in researches of diseases. In this study, we apply the concepts in complex network control to the drug identification problem. Based on these concepts, we are able to define the measurements of drug efficiency and its side effects. On the basis of the measurements, we design an algorithm to identify the drug targets. Our proposed method can also be used to verify the performance of existing drugs as well as find multiple drug combinations for a certain disease by checking whether the targets of existing drugs can make the biomolecular networks output controllable. Compared with other network based methods, our approach does not need the exact values of parameters in a network which are always unknown or hardly to be estimated accurately. In fact, our method only requires the structure of the biomolecular networks and its results are based on a general property which is the controllability of complex networks. Indeed, by applying our method to some real biomolecular networks, one can see our results are supported by existing research results.

In this study, the algorithm is designed to find the minimum number of steering nodes while it does not focus on $GDCS$. The future study will make a balance between finding a small set of steering nodes and making corresponding $GDCS$ small. In addition, we will study the controllability of biomolecular networks with specific biological information or constraints added to the networks. The drug target identification will base on the controllability of biomolecular networks instead of general complex networks.

4.5 Appendix

Proof: the number of nodes q matched to nodes in $V_{U'}$ in a maximum weight complete matching is minimum.

By contradictory. Assume the score of a maximum weight matching is $f + g \times \varepsilon$, where f and g are integers and $g * \varepsilon < 1$. In the constructed bipartite graph, the score of a complete matching is the sum of all the weights of edges matched to the nodes in V_R. For $v_{s_i} \in V_O (i = 1, ..., k)$, the weights of edges $w_{r_{s_i} u'_i}$ between $r_{s_i} \in V_R$ and $u'_i \in V'_U$ in the bipartite graph are $\varepsilon \times rcc_{s_i}$, which can be considered as $1 + \varepsilon \times rcc_{s_i} - 1$. Then every node $v_i \in V_O$ contributes a weight with integer part 1 to the matching score and every node in V'_U which has been matched contributes a weight with integer part -1, which suggests $f = p - q$.

If q nodes in V_R have matched the nodes in V'_U in the bipartite graph, the score of the maximum weight matching is

$$(p - q) + g \times \varepsilon.$$

If there is a maximum weight matching that has fewer nodes in V_R match nodes in V'_U, we denote the number of nodes as $q - q'$, where q' is a positive integer number. The score of the maximum-weight matching will be in form of

$$(p - q + q') + g' * \varepsilon,$$

where $g' \times \varepsilon < 1$.

This score is obviously larger than $f + g \times \varepsilon$, which contradicts with the fact that $f + g \times \varepsilon$ is the score of a maximum matching. So q is the minimum number of matched nodes in $V_{U'}$ in a maximum weight complete matching.

Acknowledgment

This research is supported by Natural Sciences and Engineering Research Council of Canada (NSERC), the National Natural Science Foundation of China under Grant No.61428209, and Chinese Scholarship Council (CSC).

5 Biomolecular network controllability with drug binding information

Published as: L. Wu, L. Tang, M. Li, J. Wang, and F.-X Wu, "Biomolecular network controllability with drug binding information," *IEEE Transactions on NanoBioscience*, vol. 16, no. 5, pp. 326-332, 2017.

In previous chapter, we investigated the output controllability of biomolecular networks, which considers controllability of subsets of nodes from a view of practical application. Besides the interests of controlling subsets of nodes in networks, other realistic issues should be considered in controlling biomolecular networks. Guaranteeing controllability is the first step for actually controlling a network. After identifying the steering nodes by the proposed algorithms, the following procedure can determine the input control signals which are going to be applied to the steering nodes. To actuate nodes in biomolecular networks, the input control signals are usually chemical molecules such as drugs. However, it is possible that no drugs can actuate the identified steering nodes. Since the steering node sets guaranteeing the controllability of networks are not unique, a method is required to identify specific steering node sets in biomolecular networks such that the steering nodes have more opportunities to bind to drugs.

In this chapter, an algorithm is developed to identify an MSS with pre-defined preference such that the preference values of nodes in the identified MSS is the highest among all possible MSSs of the network. To identify steering nodes with drug binding preference, three schemes to determine preference values of nodes based on their ability to bind to drugs are compared. Applications to the intracellular signal transduction network and the colitis-associated colon cancer (CAC) network show that the identified steering nodes are more likely to be known drug targets and likely to bind to more drugs. This chapter accomplishes Objective 4 of this thesis.

Abstract

Complex networks are ubiquitous in nature. In biological systems, biomolecules interact with each other to form biomolecular networks, which determine the cellular behaviors of living organisms. Controlling the cellular behaviors by regulating certain biomolecules in the network is one of the most important problems in systems biology. Recently, the connections between biological networks and structural control theory have been explored, uncovering some interesting biological phenomena. Some researchers have paid attention to

the structural controllability of networks in the context of the minimum steering sets (MSSs). However, because the MSSs for complex networks are not unique and the importance of different MSSs is diverse in real applications, MSSs with certain meanings should be studied. In this study, we investigate the MSSs of biomolecular networks by considering drug binding information. The biomolecules in the MSSs with binding preference are enriched with known drug targets and are likely to have more chemical-binding opportunities with existing drugs compared with randomly chosen MSSs, suggesting novel applications for drug target identification and drug repositioning.

5.1 Introduction

Biomolecules do not perform their functions in isolation. Diverse cellular behaviors of living organisms are the results of complicated interactions between different biomolecules. The last decade has witnessed an exceptional development of high throughput technologies, which pave the ways to the reconstruction of different types of biomolecular networks, such as gene regulatory networks, signal transduction networks and metabolic networks [155]. Because of the interactions among biomolecules in a biomolecular network, perturbing some biomolecules can affect others, which may cause the state transition of the whole network and finally change the cellular behavior. Therefore, controlling the biomolecular networks becomes an attractive research topic.

To control the biomolecular networks, the prerequisite is the modeling of the dynamics of biomolecular networks. Several dynamic models have been applied to investigate the dynamics of biomolecular networks [15, 16, 115, 156, 157]. Based on the kinetics in biochemistry, Yang *et al.* [115] constructed an arachidonic acid metabolic network (AAnetwork). The dynamics of the AAnetwork was expressed by the Michaelis-Menten equations. In their study, the authors both identified the optimal drug targets for anti-inflammation drugs and found the mechanisms of the side effects of existing drugs NSAIDs and Vioxx. Zhang *et al.* [156] proposed a p53 signaling network model and discovered a new mechanism for p53 dynamics and cell fate decision. Besides biochemical kinetics, Boolean dynamics is also widely used to describe the dynamics of biomolecular networks. Helikar *et al.* [157] supposed that the intracellular signaling transduction networks can be considered as information processing networks. To verify this hypothesis, they created a large scale network of human cell with logical mechanism of each node and concluded that the intracellular signaling transduction networks had the characteristics of nontrivial decision-making systems. Flux balance analysis [158], which is mostly used in metabolic networks, provides another perspective to represent the dynamics of biomolecular networks. Based on the flux balance analysis, several methods were developed to identify drug targets in metabolic networks [15, 16].

Many recent studies have focused on the controllability of biomolecular networks based on the linear dynamic model, producing interesting results [56, 62, 70, 87, 100, 102, 103]. For example, Wu *et al.* [56] studied the structural transittability of complex networks and identified the steering kernels for transiting phenotypes

of regulatory biomolecular networks. Wu *et al.* [70] proposed a method for drug target identification based on the output controllability of biomolecular networks. Some studies have investigated the controllability of biomolecular networks in the notion of the minimum driver node sets (MDSs) [19]. Badhwar recently *et al.* [87] investigated the phenotypic properties and the genetic correlations of the neurons which act as driver nodes in neuronal network of C.elegans. Vinayagam *et al.* [103] classified the proteins in the directed human PPI network as indispensable, dispensable or neutral, which correlate to increasing, no effect, or decreasing the cardinality of the MDS of the network by removing the proteins and edges that connect to the proteins. They found that the indispensable proteins in the human PPI network are enriched in human virus targets and drug targets, both of which provide a novel connection between the network control properties and the biological observations. However, applying independent control signals on the MDSs is a necessary condition, but not a sufficient condition, for completely structurally controlling networks. Therefore, Wu *et al.* [62] developed an algorithm to identify the minimum steering sets (MSS) of networks and compared the MSSs and MDSs of several biomolecular networks. An MSS is a minimum set of nodes required to be actuated by input control signals to have a network structurally controllable.

Studies of MDSs and MSSs provide promising insights for exploring biomolecular networks. However, since the MDSs or MSSs of a network are not unique, the algorithms for identifying MSSs or MDSs do not result in a unique set of MSS or MDS. To address this problem, Jia *et al.* classified a node as critical, intermittent or redundant if it acts as a driver node in all, some or none of all MDSs, respectively [159]. In related research, Jia *et al.* proposed a concept called control capacity, which quantifies the likelihood that a node is a driver node in an arbitrary MDS [101]. Based on this idea, Liu *et al.* classified the metabolites in a human liver metabolic network [102] and calculated the control capacities of proteins in a human signaling network [100]. Liu's studies investigated the roles of metabolites in different categories and the proteins with different control capacities.

In this study, we consider the problem of the non-uniqueness of MSSs from a practical perspective: that is, to find the MSS of biomolecular networks that should have the most chemical-binding opportunities with existing drugs. To address this issue, we take the drug binding information into consideration. Specifically we propose a method to identify the MSSs of biomolecular networks with binding preference such that the biomolecules in identified MSSs are approved drug targets or have the strong ability to bind to existing drugs. Actually, the investigation of MSSs with binding preference can be viewed as a network control problem under constraints in realistic applications. The proposed method can be used to study the control of biomolecular networks by considering both network dynamics and practical applications, which provides a novel perspective to explore biomolecular networks.

The rest of this paper is organized as follows: Section II introduces some basic concepts of structural controllability and presents the algorithm for identifying an MSS with drug binding information. The materials and the schemes for determining the preference values are also proposed in this section. Section III illustrates two applications on the intracellular signal transduction network and the Colitis-associated colon

cancer (CAC) network. Finally, Section IV concludes this study and points out some directions for future work.

5.2 Methods and materials

In this section, an algorithm is proposed to identify the MSS of a network with steering node preference after some basic concepts and results are introduced. Nodes in the MSS identified by the proposed algorithm have the maximum average preference value compared to nodes in other possible MSSs of the network. The preference values of nodes are defined based on the applications. In this section, the materials that used to determine the preference values based on drug-protein interactions are also introduced.

5.2.1 Network dynamic model and structural controllability

In this study, we describe the dynamics of a biomolecular network by the linear time-invariant nodal dynamic model, which can be described by the following equation:

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t) \tag{5.1}$$

where $\mathbf{x}(t) = (x_1(t), ..., x_n(t))^T$ is a state vector that describes the states of nodes in the biomolecular network. For biomolecular networks, states of nodes can be concentrations of metabolites or enzymes in a metabolic network or can be expression levels of genes in a gene regulatory network. A is an $n \times n$ state transition matrix. Each entry a_{ij} $(i \neq j)$ in matrix A indicates the strength of influence from biomolecule j to biomolecule i in the biomolecular network and a_{ii} is the sum of strength of self regulation and intrinsic dynamics, such as degradation, of node i. $\mathbf{u}(t) = (u_1(t), ..., u_m(t))^T$ is an input vector of m independent input control signals. From the perspective of the biomolecular networks, the input control signals can be the stimuli from environment or drugs. The $n \times m$ matrix B is an input matrix that indicates the nodes which are directly actuated by input control signals. The dynamics of a network described by equation (5.1) is denoted as system (A, B).

Because for most biological systems, it is feasible to qualify whether there is a regulatory relationship between two biomolecules, but difficult to quantify the strength of regulation, in this study, we consider the structural controllability of biomolecular networks [20]. System (A, B) is called a structural system when the entries in matrices (A, B) are either fixed zero or independent parameters.

According to the Kalman's controllability rank condition, system (A, B) is completely controllable if and only if the $n \times nm$ controllability matrix

$$\mathfrak{C} = \begin{bmatrix} B & AB & A^2B & ... & A^{n-1}B \end{bmatrix} \tag{5.2}$$

has full rank of n [34]. A structural system (A, B) is completely structurally controllable if it is possible to choose the values for the independent entries in matrices A and B such that the Kalman's controllability rank condition is satisfied [20].

To investigate the structural controllability, every system (A, B) has a corresponding graph representation G(A,B) (See Fig. 5.1). G(A,B) is a digraph which contains a set of nodes $V_A \bigcup V_U$, where $V_A = \{v_1, ..., v_n\}$ and $V_U = \{u_1, ..., u_m\}$ and a set of edges $v_j \to v_i$ for $a_{ij} \neq 0$ and $u_j \to v_i$ for $b_{ij} \neq 0$. G(A) is a subgraph of G(A, B) induced by the node set V_A. The nodes in V_A represent state nodes in the network, which corresponds to the biomolecules in a biomolecular network. The edges between nodes in V_A are indicated by the state transition matrix A, which correspond to the interactions between biomolecules. The nodes in V_U represent input nodes, which correspond to external inputs to the biomolecular network. Each node u_i in V_U of G(A,B) corresponds to the input control signal $u_i(t)$ in $\mathbf{u}(t)$. Edges from a node u_i in V_U to nodes in V_A are indicated by the ith column in the input matrix B.

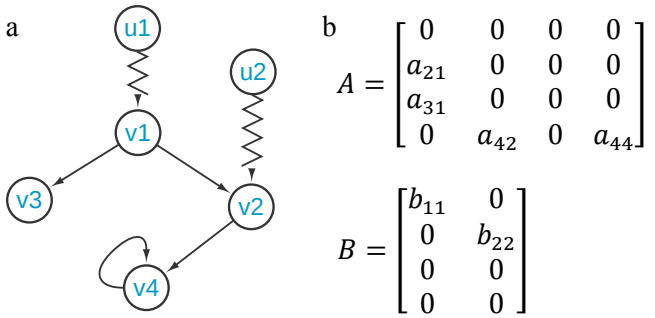


Figure 5.1: Graph representation of a system. (a): G(A,B) corresponds to system (A, B). (b): The state transition matrix and input matrix of the system (A, B).

The graph-theoretic conditions for structural controllability have been developed in previous studies [20, 120]. Before introducing structural controllability theorem, we introduce two following concepts.

***Definition* 5.1** (Inaccessibility [20, 62])**.** A node in V_A is called accessible if and only if there exists a directed paths reaching the node from an input vertex in V_U, otherwise the node is inaccessible.

***Definition* 5.2** (Dilation [20, 62])**.** The digraph G(A,B) contains a dilation if and only if there is a subset S of V_A such that the cardinality of set T(S) is smaller than the cardinality of set S, where $T(S) = \{v_j \mid (v_j \to v_i) \in E(G),\ v_i \in S\}$ and E(G) is the edge set of G(A,B). The input nodes are not allowed to belong to S but may belong to T(S).

***Theorem* 5.1** (Structural controllability theorem [20, 62])**.** A structural system (A, B) is completely structurally controllable if and only if:

i) the digraph G(A,B) contains no dilation.

ii) no node is inaccessible in V_A.

5.2.2 Algorithm for identifying MSS with steering node preference

If each node of an MSS is actuated by an independent input control signal, the resulting system G(A,B) is completely structurally controllable because both conditions i) and ii) of Theorem 5.1 are satisfied. In this

study, we develop an algorithm to identify an MSS with steering node preference of a network represented by its structure matrix A. The average of preference values of nodes in the identified MSS is the maximum among all possible MSSs of the network.

We formulate the identification of the MSS with steering node preference as a minimum cost maximum flow problem [124] in a corresponding network G_f of a given G(A) with n nodes. To construct the network G_f, firstly we divide G(A) into strongly connected components (SCCs) [125]. By contracting each SCC to a single node, the resulting graph is a directed acyclic graph. The SCCs that correspond to nodes without any incoming edge in the resulting graph are called source SCCs. Then we normalize the preference values of nodes to the range between 0 to 1. The normalized preference values are denoted as $\{PV_1, \cdots, PV_n\}$. Then G_f can be constructed in the following steps:

1. Construct a bipartite graph which contains node sets $R = \{r_1, \cdots, r_n\}$ and $C = \{c_1, \cdots, c_n\}$. Nodes r_i and c_i correspond to the node i of G(A). Connect r_j and c_i if there is a directed edge from node i to node j in G(A). The capacity and cost of all these edges in this bipartite graph are one and zero, respectively.

2. Add a sink node t to the bipartite graph and add edges from all nodes in C to the sink node. The capacity and cost of all these edges are one and zero, respectively.

3. Add a source node s to the bipartite graph.

4. Let $SCC1, SCC2, ...,$ denote different source SCCs of G(A). For each $SCCi$ that consists of more than one node, denoting its nodes as $SCCi\_1, \cdots, SCCi\_j$ $(j > 1)$:

   (a) Create two auxiliary nodes a_{SCCi_1} and a_{SCCi_2}.

   (b) Add an edge from the source to a_{SCCi_1} with the capacity and cost of the edge being one and zero, respectively. Add an edge from a_{SCCi_1} to a_{SCCi_2} with the capacity and cost of the edge being one and $n - max\{PV_{SCCi\_1}, \cdots, PV_{SCCi\_j}\}$, respectively. Add an edge from the source to a_{SCCi_2} with the capacity and cost of the edge being $j - 1$ and zero, respectively.

   (c) Add edges from a_{SCCi_2} to each node of $\{r_{SCCi\_1}, \cdots, r_{SCCi\_j}\}$ with the capacity and cost of the edges being one and $PV_{SCCi\_k}, k = 1, \ldots, j$, respectively.

5. For each source $SCCi$ that consists of only one node $SCCi\_1$, add an edge from the source node to r_{SCCi_1} with the capacity and cost of the edge being one and n, respectively.

6. For non-source $SCCi$ that consists of node $SCCi\_1, \cdots, SCCi\_j$, add edges from the source node to each node of $\{r_{SCCi\_1}, \cdots, r_{SCCi\_j}\}$ with the capacity and cost of the edges being one and $PV_{SCCi\_k}, k = 1, \ldots, j$, respectively.

Then the MSS with preference can be determined based on the following theorem [74]:

***Theorem* 5.2.** For the minimum cost maximum flow f in G_f, the MSS of the network with steering node preference consists of two types of nodes:

i) nodes in network whose corresponding nodes in R of G_f without the flow f passing through.

ii) If a source SCC_i whose all corresponding nodes in R have flow passing through, choose a node in $SCCi$ with the maximum preference value as a steering node.

Fig. 5.2 is an illustrative example of identifying the MSS with preference. The network has 4 possible sets of MSSs. The sum of preference values of nodes in the MSS $\{v_3, v_4\}$ identified by the method is the maximum compared to other MSSs (See Table 5.1).

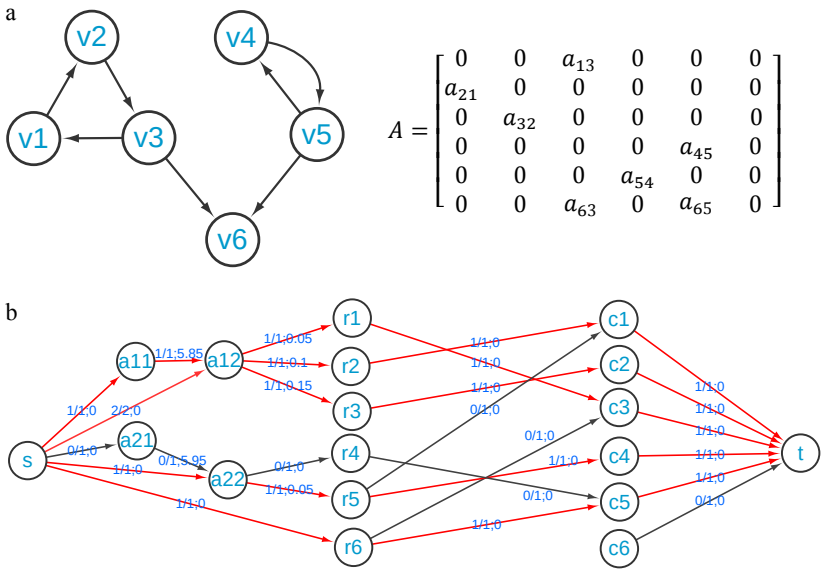


Figure 5.2: Identification of MSS with the maximum preference values. (a): A network G(A) and its corresponding transition matrix A. (b): Considering the preference, the minimum cost of maximum flow in the corresponding G_f of G(A). The labels represent (flow / capacity; cost) of the edges. The preference values of nodes v_1 to v_6 are 0.05, 0.1, 0.15, 0, 0.05 and 0, respectively. The red edges indicate the minimum cost maximum flow. The MSS with the maximum preference values is $\{v_3, v_4\}$.

MSS	$\{v_1, v_4\}$	$\{v_1, v_5\}$	$\{v_2, v_4\}$	$\{v_3, v_4\}$
Average value	0.025	0.05	0.05	0.075

Table 5.1: All possible MSSs and their average of preference values.

5.2.3 Preference values and materials

Since we focus on controlling biomolecular networks via the binding of chemicals (drugs) to biomolecules (proteins) in this study, we take the drug-protein binding information into consideration in order to have the identified steering nodes (proteins) in MSSs to be more feasibly regulated by drugs. We explore three

different schemes for defining the preference values of proteins based on the drug-protein interactions. Scheme I is based on whether a protein is a target of approved drugs: the preference value of the protein is 1 if it is a target of a drug otherwise it is 0. The preference values based on Scheme I ensures that the proteins in the MSSs with preference are enriched with drug targets. Scheme II is based on the number of drugs that can bind to the proteins. The preference values of proteins are calculated by the ratio of the number of drugs that a protein can bind with to the maximum numbers among all proteins under consideration, which range from 0 to 1. The preference values based on Scheme II ensures that the proteins in the MSSs with the preference are likely to interact with more drugs, which suggests there are more options when choosing drugs for a certain control objective. Scheme III is a combination of Scheme I and Scheme II: the preference value of a protein is the average of preference values based on Scheme I and Scheme II.

For Scheme I, 2,251 approved drugs are retrieved from DrugBank database (https://www.drugbank.ca) [160]. The proteins which are the targets of the 2,251 drugs are assigned the preference value of 1. For Scheme II, the drug-protein interaction data are acquired from STITCH database (http://stitch.embl.de) [161], which collects the confidence of chemical-protein bindings and chemical-chemical bindings. In the study, the binding information of 8,037,386 small molecule-protein pairs is downloaded from STITCH. To get the drug-protein binding information, we firstly retrieve PubChem Compound IDs [162] of all 1,652 approved small molecule drugs from DrugBank. Then we get the binding information of 295,919 drug-protein pairs from the small molecule-protein pairs based on PubChem Compound IDs of drugs. Note that Scheme II only considers the small molecule drugs due to the lack of binding data between biotach drugs and proteins. To get the binding data from STITCH, the protein names are mapped to STITCH protein IDs based on Uniprot database (http://www.uniprot.org) [163]. From STITCH, the confidence values of drug binding to a protein ranges from 0 to 1. The larger the confidence value of a drug and a protein, the more possible the drug binds to the protein. For Scheme II and Scheme III, the number of drugs that each protein can bind to is defined as the number of drugs whose binding confidence values with the protein are larger than a certain threshold s_T. In the study, we empirically choose 0.95 as the threshold s_T, which suggests that the drug-protein interactions are with high confidence.

5.3 Applications

We investigate the MSSs with chemical binding preference of two regulatory biomolecular networks. We demonstrate that the proposed method is capable to identify the MSSs of networks such that the steering nodes in the identified MSSs are more likely to be regulated by existing drugs, which supports the applicability of the method.

5.3.1 Intracellular signal transduction network

Intracellular signal transduction is a process that chemical signals transmit from the outside of cells to cellular systems, such as the nucleus or cytoskeleton, which could generate appropriate responses to those signals. To study this process, an intracellular signal transduction network is created by Helikar *et al.* [157], which consists of 130 nodes and 558 edges (See Fig. 5.3). The majority of nodes represent proteins and the other nodes represent metabolites in the network. The edges indicate the regulatory relationships between nodes. In the network, there are three major receptor families, which are receptor tyrosine kinases (RTKs), G protein-coupled receptors (GPCRs) and Integrins. In addition, proteins in the network such as Akt, Erk, Rac, Cdc42 and SAPK are related to cell apoptosis, gene transcription, cytoskeletal regulation and stress of the cell, respectively. Though the model is a nonspecific network and it does not represent any specific cell type, the nodes and interactions in the network can be found in a wide range of cell types. Therefore, the network is a general model that is feasible for investigating the common signal transduction in cells.

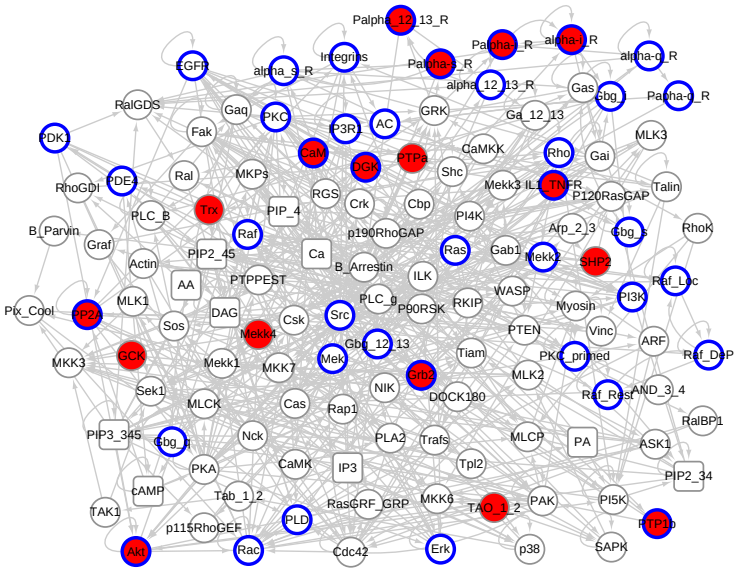


Figure 5.3: Intracellular signal transduction network. The nodes represented by circles correspond to proteins and the nodes represented by rectangles correspond to metabolites in the network. The drug target proteins are indicated by the blue circles. The red nodes corresponds to the MSS based on the preference Scheme III.

The cardinality of an MSS of this network is 17, which suggests that to completely control the network, there are 13.1% of nodes in the network should be actuated by input control signals. To identify biologically meaningful MSSs, we investigate the MSSs with drug binding preference based on the three schemes that are defined in the previous section. For each node that represents a group of proteins, we calculate the number of drugs that can bind to any one protein in the group. In this network, there are 10 nodes having more than 10 binding drugs and 9 of them corresponds to receptor proteins. Because the maximum number of binding drugs of a node is 109, in order to attenuate the influence of outliers on the normalized preference values,

we consider the number of drugs that a node can be bound to as 10 if the number is more than 10 for this network.

The comparison of different MSSs are shown in Table 5.2. The first row indicates the number of steering proteins in the MSSs which are drug targets. The second row indicates the average number of drugs that each protein in the MSSs can be bound to. The four columns in Table 5.2 show the number of drug targets and the average number of drugs that a protein is likely bound to in the MSSs based on Scheme I, Scheme II, Scheme III and the average of random selections, respectively. The random selection is based on 1000 runs of the algorithm developed in our previous study [62] for identifying MSS without preference: to get multiple MSSs for different runs, the nodes are randomly relabeled in each run. The results with Scheme I indicate that there are at most 11 proteins in the identified MSS can be drug targets while each steering node has a small number of binding drugs. Compared to the random MSSs, the MSSs with preference Scheme I is significantly enriched with drug targets. The results with Scheme II indicate that while the number of drug targets in it are smaller than the MSS with Scheme I, each protein in the identified MSS can be bound by 3.71 drugs averagely, which is significantly larger than the average number that each protein in a random MSS can interact with. The different results of MSSs with Scheme I and Scheme II suggest that Scheme I and Scheme II capture different aspects of drug binding preference respectively. Therefore, it is reasonable to investigate Scheme III, which is a combination of Scheme I and Scheme II. As expected, the MSS with Scheme III is enriched with both drug targets and interactions with drugs, which satisfies our objective that the identified MSSs should feasibly be controlled by drugs in real applications.

	Scheme I	Scheme II	Scheme III	Random selection
Drug targets	**11**	8	**11**	4.32
Interactions	2.76	**3.94**	**3.71**	1.08

Table 5.2: Interactions between MSSs and drugs. The first row shows the number of proteins which are targets of approved drugs in the MSSs with different preference scheme. The second row shows the average number of drugs that each protein in different MSSs likely interacts with.

Proteins in the MSSs with Scheme III are highlighted as red nodes in Fig. 5.3. In this MSS, the alpha-i_R, Palpha-i_R, Palpha-s_R and Palpha-12_13_R have the highest preference values, which are both GPCRs. The observation is in agreement the fact that receptor proteins, especially the GPCRs which are the targets of 50% of drugs, are enriched with drug targets [164, 165]. Therefore, it is reasonable and applicable to control the network by regulating the receptor proteins. Other proteins in the MSS also play important roles in regulating the cellular behaviors. For example, RAC-alpha serine/threonine-protein kinase (AKT) also has the highest preference value in the network. AKT is a critical regulator of many processes such as metabolism, proliferation, cell survival, growth and angiogenesis [166–168]. In addition, AKT is also also found to be critical in the tumor development. For example, inhibitors of Akt have been investigated to treat

cancers such as neuroblastoma [169] and Arsenic trioxide, which is an inducer of AKT, has been used to the treatment of acute promyelocytic leukemia (APL) [170].

5.3.2 CAC network

It has been discovered that the inflammation and cancer are closely related [171]. To deeply understand the mechanism of inflammation-associated cancer, Lu *et al.* constructed a CAC network by integrating the extracellular microenvironment and intracellular signalling pathways [172], which contains 70 nodes and 154 regulatory interactions. Based on biological functions, the nodes in the networks can be divided into four groups: extracellular immune microenvironment, inflammatory signalling, cell proliferation and apoptosis.

In the previous application, we show that Scheme III outperforms the other two schemes. Therefore, we only consider preference with Scheme III in this application. Applying our method to the CAC network, we identify the MSSs with preference Scheme III, which is supposed to identify the most feasible MSS to completely control the network. There are 6 steering nodes in the identified MSS. Four steering proteins in the MSS are drug targets and each protein can interact with 4.17 drugs averagely. We also investigate 1000 randomly chosen MSSs without preference, in which averagely 2.14 steering proteins are drug targets and the number of small molecule drugs that each steering protein can interact with is 1.87. Compared to the randomly chosen MSSs, the proteins in the MSS with preference are significantly enriched with drug targets and interactions to drugs.

The 6 steering proteins are IL6, TNFA, CCL2, CYTC, APC and SOD, out of which IL6, TNFA and CCL2 are related to the extracellular immune microenvironment, CYTC2 is related to the apoptosis and APC and SOD are related to the cell proliferaction process. It can be observed that the steering nodes belong to different function groups, which is consistent with the knowledge that cancer is a complex disease and drugs should be applied to multiple targets for cancer therapy [173]. In the identified MSS, APC is a steering nodes that appears in all the possible MSSs of the CAC network. In fact, researchers have discovered that the colon cancer may be caused by mutations of the APC gene [174], which suggests the importance of regulating the APC for controlling the CAC network. IL6 is the target of drug Siltuximab, which has been investigated for the treatment of several types of cancers [175, 176]. CCL2 is an important mediator of the cell migration and proliferation of prostate cancer [177] and breast cancer [178]. TNFA, which is produced mainly by macrophages, is a cell death factor [179]. CYTC plays a important role in apoptosis. When CYTC is released into the cytosol, it would activate caspases, which are responsible for destroying the cell [180]. Though SOD is not a drug target, it can interact with 2 drugs based on STITCH database.

5.4 Conclusion

In this study, we have investigated the controllability of biomolecular networks by combining drug-protein binding information. A minimum cost maximum flow algorithm has been developed to identify an MSS based

on the preference. We applied the algorithm to the intracellular signal transduction network and the CAC network. The biomolecules in the MSSs with binding preference are enriched with known drug targets and are likely to have more interactions with drugs compared with randomly chosen MSSs with no preference. In addition, our results are supported by existing research results, which suggests that our proposed method could be a promising tool for drug target identification and drug repositioning.

In the future work, other control objectives, such as output controllability with binding information, could be explored. In addition, other meaningful node preference, such as drug specificity, could be considered for identifying meaningful MSSs of biomolecular networks.

Acknowledgment

This research is supported by Natural Sciences and Engineering Research Council of Canada (NSERC), the National Natural Science Foundation of China under Grant No. 61428209 and No. 61622213, and Chinese Scholarship Council (CSC).

6 CytoCtrlAnalyser: a Cytoscape app for biomolecular network controllability analysis

Published as: L. Wu, M. Li, J. Wang, and F.-X Wu, "CytoCtrlAnalyser: a Cytoscape app for biomolecular network controllability analysis", *Bioinformatics*, vol. 34, no. 8, pp. 1428-1430, 2017. Supplementary Information along with the published paper is included in Section 6.4 - Section 6.6.

Chapter 2 has introduced various methods to investigate controllability of complex networks and Chapters 3, 4 and 5 have proposed three algorithms for identifying steering node sets. These methods have provided powerful tools for understanding biomolecular networks as well as general complex networks. However, the developed algorithms are distributed in the literature and there is no easy access for using the algorithms.

In this chapter, a software system called CytoCtrlAnalyser is developed, which integrates nine network controllability algorithms including ours and others in the literature. A user guide for CytoCtrlAnalyser is given which introduces the integrated algorithms and presents two applications to biomolecular networks. This chapter also includes the details of implementation of the algorithms in CytoCtrlAnalyser and basic knowledge of network controllability theorems. Objective 5 of the thesis is accomplished in this chapter.

Abstract

Summary: Studying the controllability of biomolecular networks can result in profound knowledge about molecular biological systems. However, there is no comprehensive and easy-to-use platform for analyzing controllability of biomolecular networks although various algorithms for analyzing complex network controllability have been proposed recently. In this application note, we develop CytoCtrlAnalyser which is a Cytoscape app to provide a comprehensive platform for analyzing controllability of biomolecular networks. Nine algorithms have been integrated in CytoCtrlAnalyser. With network topologies and customized control settings imported into CytoCtrlAnalyser, users can identify the steering nodes which should be actuated by input control signals for achieving different control objectives as well as investigate the importance of nodes from different perspectives in the controllability of networks. CytoCtrlAnalyser offers a tool for many promising applications, such as identification of potential drug targets or biologically important nodes in biomolecular networks.

Availability and implementation: Freely available for downloading at `http://apps.cytoscape.org/`

`apps/cytoctrlanalyser`.

6.1 Introduction

With development of high throughput technology, methods are proposed to construct biomolecular networks from omics data [31, 181]. Biomolecular networks are widely used to model, analyze and understand the mechanisms of biological processes. However, the ultimate goal is to change the behaviors or states of biomolecular networks to the desired ones. Recently, Liu *et al.* [19] firstly investigated the controllability of complex networks based on the structural controllability theorem [20]. Since then, several researchers have expanded the network controllability studies and their methods have been applied to studying biomolecular networks [56, 62, 67, 70, 74, 92, 100, 103]. Many interesting biological phenomena have been discovered by investigating the controllability of biomolecular networks, which suggests promising applications for investigating the dynamics of biological systems in terms of network controllability.

However, algorithms for analyzing network controllability are scattered in literature and thus they are inconvenient for the researchers to use. Therefore, a comprehensive platform for running existing controllability algorithms is beneficial to researchers. Here we integrate nine recently developed algorithms to establish a comprehensive platform, called CytoCtrlAnalyser, which is a Cytoscape app that enables users to conveniently analyze various types of controllability of biomolecular networks within the Cytoscape environment [28]. We would like to mention that CytoCtrlAnalyser can be applied to other complex networks although this study mainly focuses on biomolecular networks. To our best knowledge, CytoCtrlAnalyser is the first platform that offers such comprehensive calculations and integrates algorithms for investigating the controllability of biomolecular networks (as well as other complex networks).

6.2 Description of CytoCtrlAnalyser

Functions: Based on the Cytoscape environment, CytoCtrlAnalyser provides a user friendly and straightforward interface (Fig. 6.1a). CytoCtrlAnalyser integrates nine recently developed algorithms for analyzing complex network controllability with different objectives (Fig. 6.1b). The developed algorithms investigate the network controllability from two aspects. On the one hand, some algorithms identify nodes which should be actuated by control signals for different control objectives. From this aspect, CytoCtrlAnalyser integrates algorithms for identification of Minimum Driver node Set (MDS) [19], Minimum Steering node Set (MSS) [62], MSS with preference [74], steering nodes for transittability [56] and steering nodes for output controllability [67, 70]. On the other hand, several concepts and algorithms are studied to capture the importance of nodes from various perspectives. CytoCtrlAnalyser implements four algorithms to calculate control centrality [182], control capacity [101], node classification [103] and probability of each node in a random MSS [92], respectively. For more details of functions, see Supplementary Information in Subsection 6.4.1.

Input: All algorithms in CytoCtrlAnalyser require a network as input. In addition, user customized data of control settings is also required for algorithms of transittability, output controllability and MSS with preference. For transittability and output controllability, each node should be assigned a Boolean value to indicate its role (e.g. an output node or not). For MSS with preference, each node should be assigned a preference value. The network topology and the user customized data can be stored in .txt format and imported by the standard Cytoscape method (Fig. 6.1c).

Output: The results of algorithms are displayed in the node table panel (Fig. 6.1d). In the node table panel, each row corresponds to a node in the network. For each algorithm selected by a user, a corresponding column will be created. The results of MDS, MSS, MSS with preference, transittability and output controllability are sets of steering nodes, respectively. The steering nodes are assigned true values while other nodes are assigned false values. The control capacity values, control centrality values, classification and probability of nodes in a random MSS are displayed in the corresponding columns.

See Supplementary Information in Section 6.5 for implementation of CytoCtrlAnalyser and Section 6.6 for network controllability theorems.

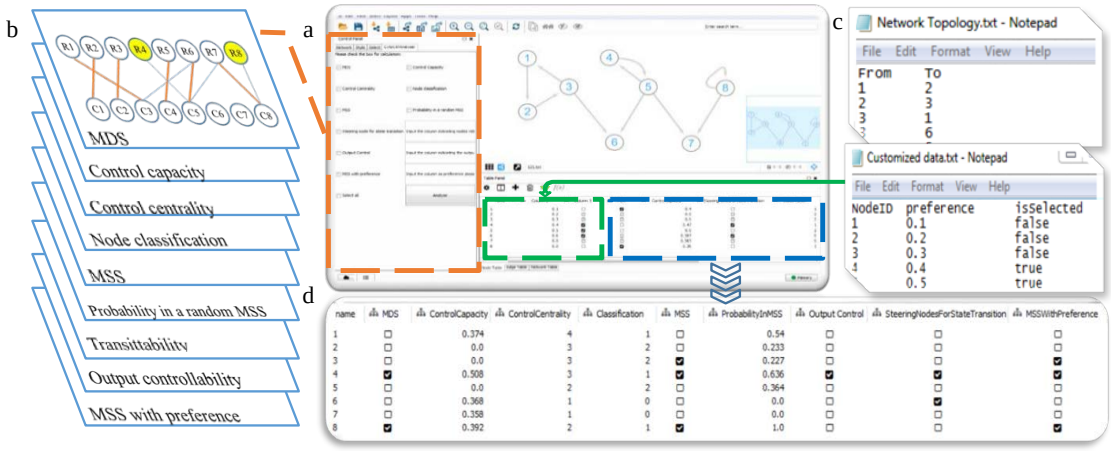


Figure 6.1: Overview of CytoCtrlAnalyser. (a): CytoCtrlAnalyser interface in Cytoscape environment. (b): Algorithms integrated in CytoCtrlAnalyser. (c): Network topology data and customized input data. (d): Display of results.

6.3 Case studies

To illustrate how CytoCtrlAnalyser can facilitate researchers to investigate the controllability of biomolecular networks, we presented two case studies where CytoCtrlAnalyser is applicable. Wu *et al.* [92] investigated the MSSs of colitis-associated colon cancer (CAC) network. Based on drug binding information, an MSS with the highest binding opportunity with drugs has been identified. Instead of implementing complicated algorithm developed in (Wu et al., 2016b) to identify the MSS with preference, researchers could conveniently get the results through three steps with CytoCtrlAnalyser. The first step is importing the CAC network and

preference values to Cytoscape. The second step is to check the checkbox of MSS with preference and indicate the name of column which stores the preference values of nodes. Third, by simply clicking the Analyze button, a corresponding column would be created in the node table and nodes in MSS with drug binding preference would be indicated by check marks. The MSS with drug binding preference identified by CytoCtrlAnalyser consists of 6 nodes, which are APC, SOD, IL6, TNFA, CCL2 and CYTC. This result is identical to the result in Wu??s study [92]. For more details of using CytoCtrlAnalyser and biological explanation of the results, see Supplementary Information in Subsection 6.4.4.

To explore the effect of removing a node to the controllability of networks, Vinayagam *et al.* [103] classified nodes as indispensable, dispensable or neutral. Proteins (nodes) in a directed human PPI network have been investigated based on this classification. With CytoCtrlAnalyser, the classification of nodes can be quickly done by importing the human PPI network to Cytoscape, checking the node classification checkbox and clicking the Analyze button. The classification result is consistent with the result in Vinayagam??s study [103]. See Supplementary Information in Subsection 6.4.4 for more details.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC); the National Natural Science Foundation of China [Grant number 61772552 and No.61622213]; and Chinese Scholarship Council (CSC).

Conflict of Interest: none declared.

6.4 User guide

6.4.1 Functions of CytoCtrlAnalyser

The CytoCtrlAnalyser implements algorithms for studying biomolecular network controllability based on nine recently proposed concepts: Minimum driver node set (MDS) [19], Minimum steering set (MSS) [62], MSS with preference [74], steering nodes for state transittability [56] and steering nodes for output controllability [67], control centrality [63], control capacity [101], classification [103] and probability of each node in a random MSS [92]. The following paragraphs give the descriptions of algorithms that are implemented for investigating network controllability.

MDS: A network is completely controllable if it can be steered from any initial states to any desired final states in finite time. The MDS [19] is a minimum set of nodes in the network which should be actuated by independent control signals such that the "no dilation" condition of structural controllability theorem [62] (See Section 6.6.2 for detail) for completely controlling a network can be satisfied. Applying independent control signals to an MDS is a necessary condition for completely controlling a network. In biomolecular networks, nodes in MDSs play crucial roles in controlling the networks. Biological functions of MDSs have

been investigated in different biological networks [87, 100]. In addition, concepts such as control capacity and node classification are defined based on MDSs of networks.

MSS: The MSS [62] is a minimum set of nodes in the network which should be actuated by control signals such that the network is completely structurally controllable. Compared to the MDS, applying independent control signals to an MSS is a sufficient and necessary condition for completely controlling a network, which satisfies all conditions of structural controllability theorem [62]. Examples are given in [62] to compare MDSs and MSSs of biomolecular networks.

MSS with preference: Since the MSSs of a network are not unique, an algorithm has been developed to identify an MSS with certain pre-defined preferences. When each node is assigned a preference value, the algorithm identifies an MSS that has the maximum average preference value among all the possible MSSs of the network [74]. In [92], MSSs with drug binding preferences have been identified, which suggests a feasible way to control biomolecular networks as well as novel applications for drug repositioning.

Transittability: In many applications, we do not need to completely control a network, but to steer the network from one specific state to another specific state. Therefore, the concept of transittability has been studied [56]. An algorithm has been designed for identifying steering nodes which should be actuated by control signals such that the network is structurally transittable between these two specific structural states. Steering nodes for state transittability of biomolecular networks are investigated in [56], suggesting that the steering nodes play important roles for state transitions of biomolecular networks.

Output controllability: In CytoCtrlAnalyser, users need to define a set of nodes which correspond to the output. The implemented algorithm can be used to identify the steering nodes which should be actuated by control signals such any output state of the system can driven to any other output state [67]. In [67, 70], output controllability has been used to identify potential drug targets in biomolecular networks such that the states of output nodes can be controlled.

Control centrality: Control centrality of a node equals to the dimension of the controllable subspace or the size of the controllable subnetwork when a control signal is actuated only on the node [63].

Control capacity: Since the MDSs of a given network is not unique, the control capacity measures the likelihood of each node appearing in a random MDS [101]. Control capacity values of nodes in a human liver metabolic network have been studied [102].

Node classification: The MDSs of a network are not unique, but the cardinality of the MDSs are the same. Therefore, nodes in a network can be classified as indispensable, neutral or dispensable, which correlate to increasing, no effect, or decreasing the cardinality of the MDS of the network by removing the node and edges that connect to the node [103]. Biological roles of different types of proteins in a directed PPI network have been studied in [103].

Probability in an MSS: Similar to the non-uniqueness of MDSs of networks, there are different MSSs of a same network. This algorithm quantifies the probability of each node appearing in a random MSS [92]. Nodes with higher probabilities in an MSS have been suggested to play more important roles in controlling

the network. In addition, by using this algorithm, we have been able to compare preference values of a randomly selected MSS and an MSS selected with preference.

6.4.2 Quick Start

Following is a short quick start for the usage of CytoCtrlAnalyser:

1. Download and install Cytoscape from `http://www.cytoscape.org/`
2. Start Cytoscape, install the CytoCtrlAnalyser.jar through App Manager (Apps → App Manager → Install App → Search *CytoCtrlAnalyser* → Install).
3. Import or open a network.
4. Click **Apps → CytoCtrlAnalyser → Open.**
5. Choose one or multiple algorithms in CytoCtrlAnalyser interface on control panel.
6. Import customized data if one of the three algorithms (MSS with preference, transittability and output controllability) is selected. **(i)** Click **File → Import → Table → File...** and select the data file. **(ii)** Use Cytoscape interface to import the user customized data from file.
7. Click **Analyze** button.

Three algorithms, which are MSS with preference, transittability and output controllability, require customized data of control settings. Each of the three algorithms requires the user to indicate a column in the node table panel. The data types are listed:

MSS with preference: Each node should be assigned a real positive number as the preference for selection of MSS. The data type could be either Integer, Long, Float or Double.

Transittability: Each node should be assigned a Boolean value which indicates whether the state of the node is going to be changed according to the control objective. Nodes with *True* values could be changed to any states in finite time while states of nodes with *False* values would remain unchanged at the end of the control process.

Output controllability: Each node should be assigned a Boolean value which indicates whether the node is corresponding to the output of the network. *True* means the node corresponds to one output of the network while *False* means the node is not the output of the network.

6.4.3 Illustrating example

Import network and open CytoCtrlAnalyser

The first step is to import a network to be studied and open CytoCtralAnalyser app. Fig. 6.2 is an example of a network with 8 nodes and 10 edges and the CytoCtrlAnalyser interface in Cytoscape.

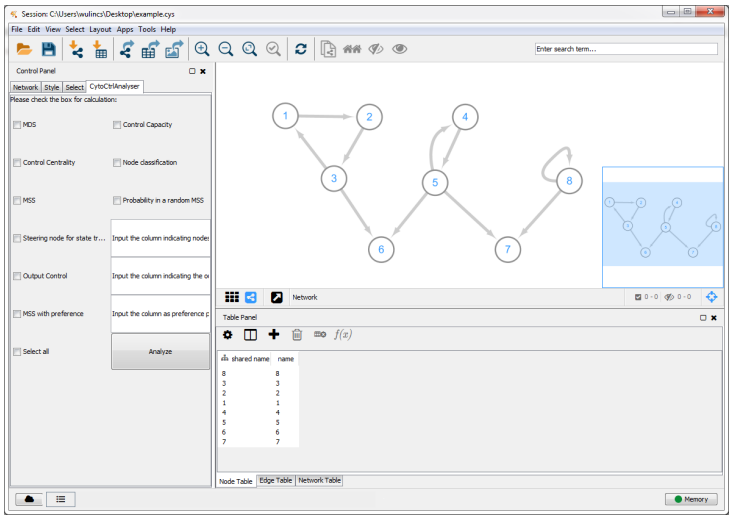


Figure 6.2: A example network and the CytoCtrlAnalyser interface.

Import customized data

The data should be stored in a table. There are two ways to manage the table which stores the customized data. The first method is to create data columns in the node table panel by using the Cytoscape interface. The second method is to import the data from existing files by using the table importing function of Cytoscape. Multiple file formats are supported, such as .txt, .xlsx and .csv. Following is an introduction for importing the customized data from files.

1. Click **File** → **Import** → **Table** → **File...** and select the data file.

2. Import the columns from file. See Fig. 6.3.

To import the customized data, a table file should be created. In Fig. 6.3a, the first column is the names of nodes in Cytoscape, which is used as the key to map the nodes and node attributes. The second column in Fig. 6.3a is the preference values of nodes for the MSS identification with preference. For CytoCtrlAnalyser, the preference values could be any positive numbers. The third column contains Boolean values for nodes, which is used as the customized data for analysing the network transittability or output controllability. In Fig. 6.3a, node 4, node 5 and node 6 are set as *True*. For network transittability, the states of node 4, node 5 and node 6 could be changed to any values at the end of control process while the states of other nodes would remain unchanged. For output controllability, the states of node 4, node 5 and node 6 could be changed to any values at the end of control process while the states of other nodes are not considered in the output controllability study.

Fig. 6.3b is the interface of importing the table file to Cytoscape. The first column is selected as the key, which would match the values to corresponding nodes in Cytoscape. Users need to indicate the data types of different columns: the second column is set as *Floating point* and the third column is set as *Boolean*. Fig.

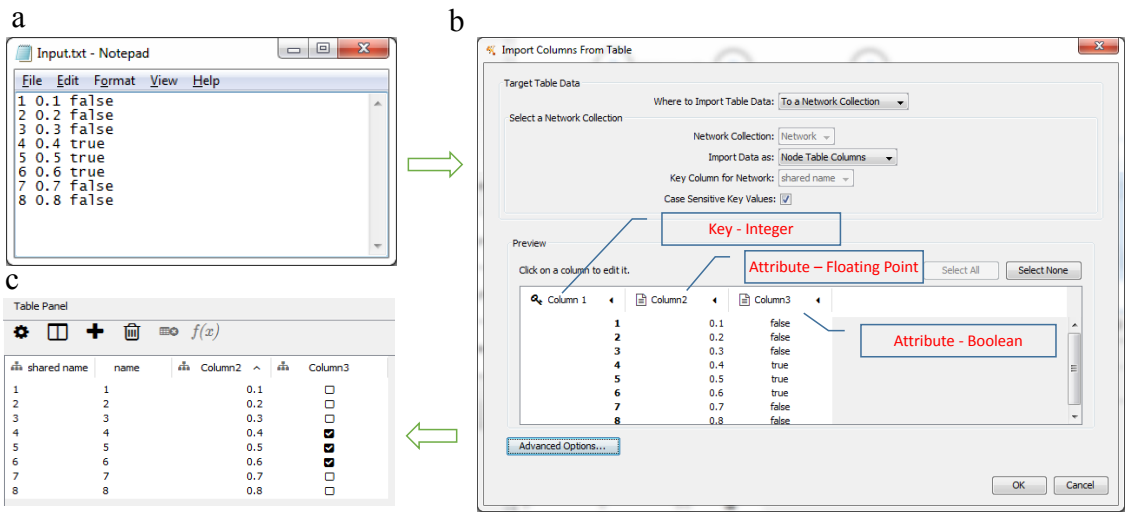


Figure 6.3: Procedures of importing customized data. (a): An .txt file with preference values and Boolean values. (b): Interface of Cytoscape for importing the data to the node table panel. (c): Imported customized data in the node table panel.

6.3c is the data imported to Cytoscape shown in node table panel: Column 2 saves the preference values of nodes and Column 3 indicates the node whose states are supposed to be changed in state transition or output control.

Analyse controllability of network

Nine algorithms for network controllability analyses are included in CytoCtrlAnalyser. To apply controllability algorithms, users simply need to check the algorithms they would like to use. For the state transittability and output controllability, the nodes whose states are going to be changed are required to be indicated. For MSS with preference, the preference value of each node should be input. To indicate the customized data, the names of columns containing the corresponding data are required to be indicated in the text boxes on the right of the algorithm check boxes.

Fig. 6.4a is an overall interface of the input and the result display. In the illustrating example, we indicate Column 2 as the user input data for the MSS with preference algorithm and Column 3 as the user input data for state transittability and output controllability. Then we check *Select all* box and press the *Analyze* button (See Fig. 6.4b).

The results are displayed in Fig. 6.4c. For the example network, node 4 and node 8 make up one possible MDS of the network. The control capacity of node 1 to node 8 are 0.374, 0, 0, 0.508, 0, 0.368, 0.358, and 0.392, respectively. The control centrality of node 1 to node 8 are 4, 3, 3, 3, 2, 1, 1 and 2, respectively. For node classification, node 1, node 4 and node 8 are classified as neutral nodes (labeled as 1); node 2, node 3 and node 5 are indispensable nodes (labeled as 2); node 6 and node 7 are dispensable nodes (labeled as 0). Node 3, node 4 and node 8 make up one MSS of the network. The possibility of node 1 to node 8 that appear in an MSS are 0.54, 0.223, 0.227, 0.636, 0.364, 0, 0 and 1, respectively. As Column 3 indicated, if we

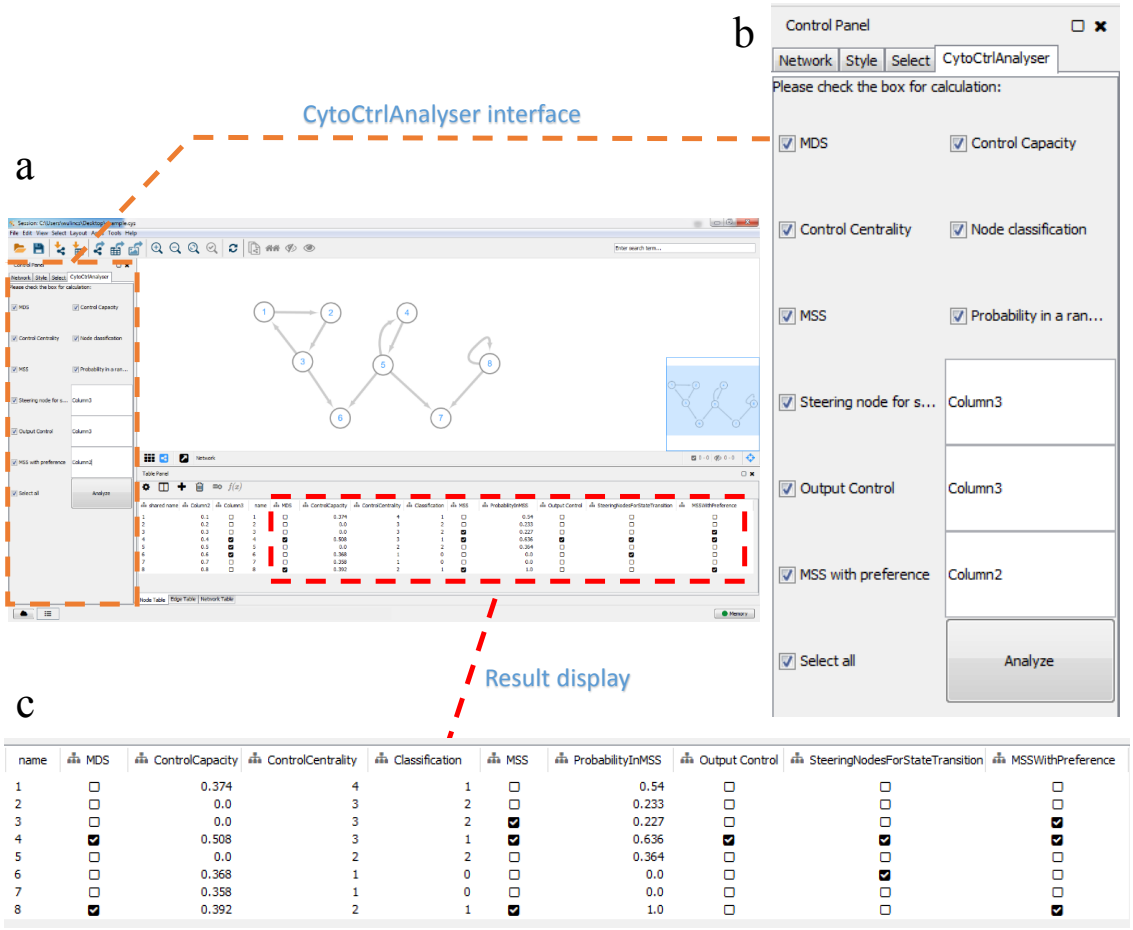


Figure 6.4: Procedures of importing customized data. (a): The .txt file which indicates the preference values for choosing the MSS and the nodes for the transittability or output controllability. (b): The interface of Cytoscape for importing the data to the node table panel. (c): The imported customized data shown in the node table panel.

would like to change the state of node 4, node 5 and node 6 without effecting the states of other nodes, node 4 and node 6 are required to be actuated by control signals. According to the output controllability, if we would like to change the states of node 4, node 5 and node 6 while not considering the states of other nodes, it can be achieved by actuating node 4 only. Based on the input preference values, the node 3, node 4 and node 8 make up the most preferred MSS.

6.4.4 Application examples

MSSs with drug binding preference of colitis-associated colon cancer (CAC) network

In [92], Wu *et al.* investigated the MSS with drug binding preference of CAC network. This subsection gives an application example to show how to get the MSS with drug binding preference by using CytoCtrlAnalyser.

1. Acquire the CAC network from paper [172]. The network file is reformed to .txt format which can be supported by Cytoscape (Fig. 6.5a).

2. Import the CAC network to Cytoscape and open the CytoCtrlAnalyser interface (Fig. 6.5b).

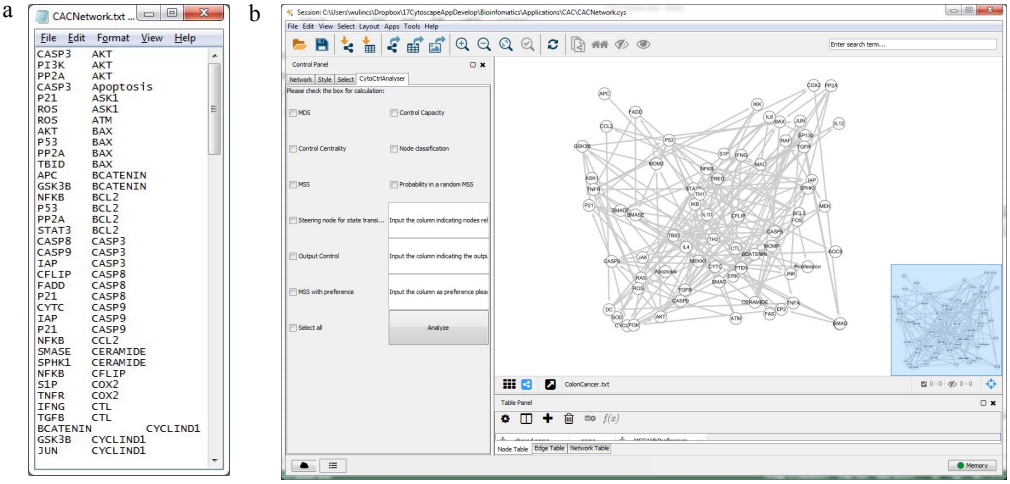


Figure 6.5: CAC network file opened in Cytoscape.

3. Check the *Probability in a random MSS* checkbox and press *Analyze* button. After results appear in node table, sort the table in descending order based on values in column *ProbabilityInMSS*. We can see that there are 20 nodes with nonzero values, which means only these 20 nodes can be steering nodes in MSSs of the CAC network (Fig. 6.6a). The reason is that the other nodes never appear in any MSSs and their preference values do not impact the result of MSS with preference.

4. Acquire drug binding preference values of 20 nodes identified in step 3 by using the strategy proposed in [92]. Because nodes CTL, DC, MAC and TREG correspond to different types of cells, which do not correspond to specific proteins, the preference values of these nodes are set to 0.

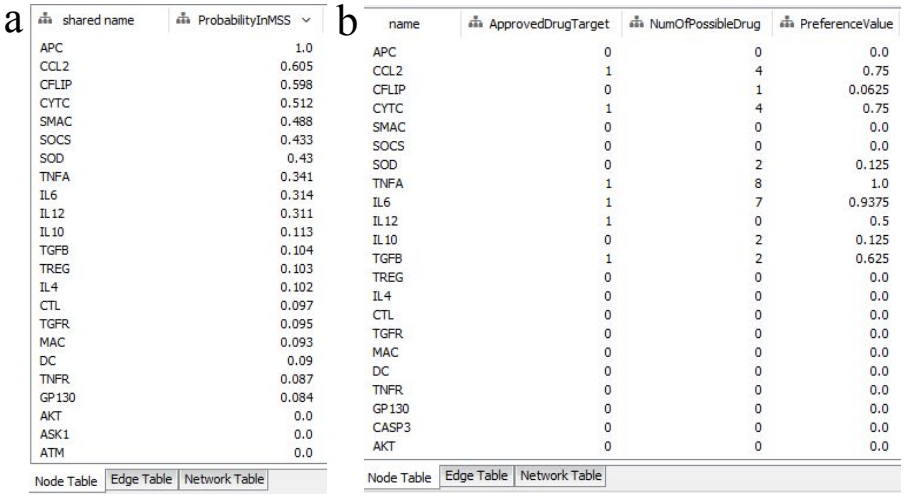


Figure 6.6: Importing preference values of nodes to Cytoscape.

5. Import the preference values to Cytoscape. In Fig. 6.6b, the first column lists the node names, the second column indicates whether the node is a target of any approved drugs and the third column shows to number of drugs that can bind to the node with high confidence according to STITCH database (http://stitch.embl.de) [161]. The MSS is identified based on the preference values in the forth column, which are calculated by values in the second and the third columns.

6. Check the checkbox *MSS with preference* and input *PreferenceValues* in the following text area (Fig. 6.7a).

7. Press *Analyze* button and the MSS with preference is indicated in column *MSSWithPreference* of node table (Fig. 6.7b).

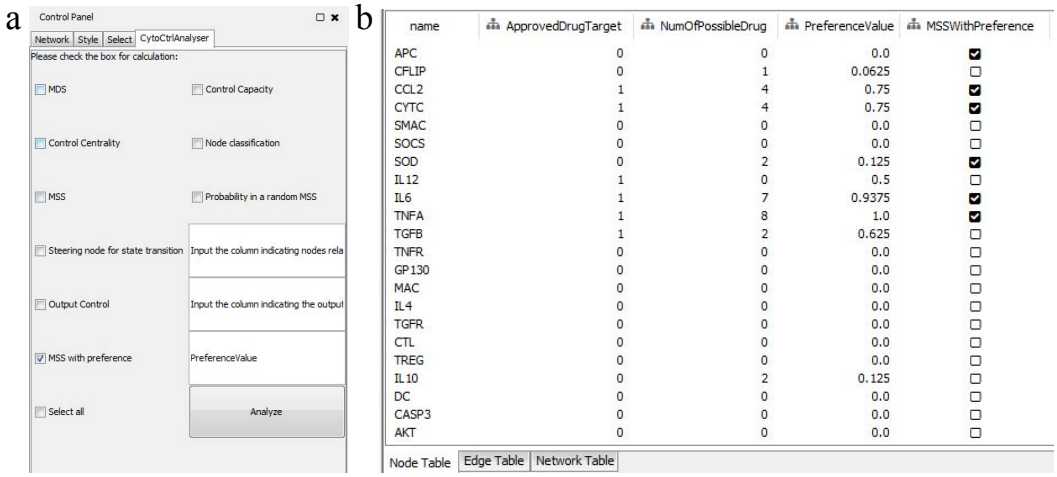


Figure 6.7: Identifying MSS with preference.

From Fig. 6.7b, we can see that the MSS with drug binding preference consists of 6 steering nodes (proteins), which are APC, CCL2, CYTC, SOD, IL6 and TNFA. The result is exactly the same as the result in [92]. According to analyses in [92], controlling the CAC network by actuating states of the 6 identified proteins is biologically meaningful. Firstly, these steering nodes belong to three different function groups, which are cell proliferation process, extracellular immune microenvironment and apoptosis. This observation is consistent with the knowledge that cancer is a complex disease and drugs should be applied to multiple targets for cancer therapy [173]. For individual node in the MSS with drug binding preference, APC is a steering nodes that appears in all the possible MSSs, which suggests the importance of APC in completely controlling CAC network. Actually, researchers have discovered that the colon cancer may be caused by mutations of the APC gene [174]. CCL2 is observed as an important mediator of the cell migration and proliferation of prostate cancer [177] and breast cancer [178]. Drug Siltuximab, which targets at IL6, has been investigated for the treatment of different cancers [175, 176]. In addition, TNFA is a cell death factor [179] and CYTC is related to apoptosis.

In the table, we can also find that 4 steering nodes in the MSS are drug targets and each node can interact with 4.17 drugs averagely. By multiplying corresponding values in column *ApprovedDrugTarget* and column *ProbabilityInMSS*, we can find that averagely there are only 2.19 steering nodes that are targets of approved drugs in a random MSS without preference. Similarly, by multiplying corresponding values in column *NumOfPossibleDrug* and column *ProbabilityInMSS*, we can find that each steering node in a random MSS can interact with only 1.84 drugs averagely. The results demonstrate that steering proteins in the MSS with drug binding preference are significantly enriched with drug targets and interactions to drugs, which suggests that controlling CAC network by actuating states of steering proteins in the MSS with drug binding preference is more feasible compared to actuating nodes in a randomly selected MSS.

Node classification of human directed PPI network

Vinayagam *et al.* [103] classified the proteins in the directed human PPI network as indispensable, neutral or dispensable, which correlate to increasing, no effect or decreasing the cardinality of MDS of the resulting network by removing the proteins and edges that connect to proteins.

CytoCtrlAnalyser implements an algorithm to classify nodes based on this classification method. To get the classification, firstly the directed human PPI network can be acquired from supplementary materials of reference [4]. Then the network is imported to Cytoscape and CytoCtrlAnalyser is opened. After this, the classification of nodes can be calculated by checking the *Node classification* checkbox and pressing the *Analyze* button. The result is shown in column *Classification* of node table panel, in which 0, 1 and 2 correspond to dispensable, neutral and indispensable, respectively (Fig. 6.8).

To verify the classification result from CytoCtrlAnalyser, we imported the classification of nodes in reference [103] to column *Node Class* of node table panel. By comparing the results, we found that for all 6,338 nodes in the network, only classification of protein PRG2 is different. PRG2 is classified as a indispensable

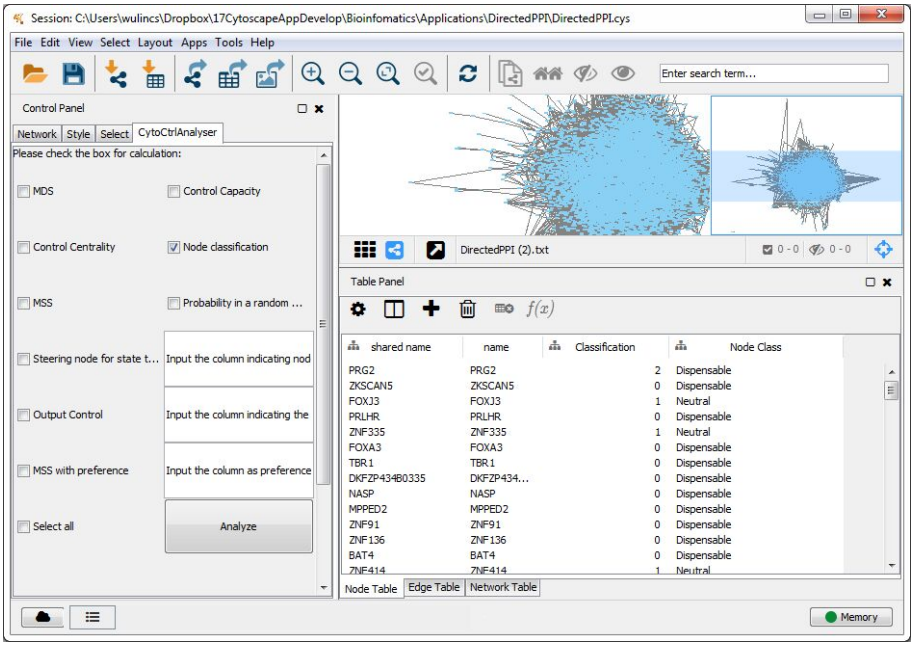


Figure 6.8: Node classification of human directed PPI network.

node by CytoCtrlAnalyser while [103] suggested it is dispensable. To verify the classification of PRG2, we used MDS algorithm in CytoCtrlAnalyser to calculate MDS of the network and MDS of resulting network after the removal of node PRG2, respectively. We found that the cardinalities of MDSs are 2,282 and 2,283, respectively. Removing node PRG2 from the network increases the cardinality of MSS, which suggests PRG2 is an indispensable node and the classification result from CytoCtrlAnalyser is correct. The reason of reference [103] has different classification of PRG2 is probably that the PPI network used in [103] is slightly different from the PPI network in [4]. It is because the network in [103] has 6,339 nodes and 34,813 directed edges while the network file downloaded from [4] has 6,337 nodes and 34,814 directed edges.

CytoCtrlAnalyser provides a platform for users to get access to different network controllability algorithms. However, biological interpretations of results from CytoCtrlAnalyser depend on specific problems, which require researchers to analyse the results based on their own knowledge and proposes. In [103], authors found that indispensable proteins in the human PPI network are enriched in human virus targets, drug targets or disease-causing mutations. Their study provides a fresh classification method based on network controllability, which shows distinct biological properties in the context of essentiality, conservation and regulation. Therefore, CytoCtrlAnalyser provides a convenient tool for future studies on different biomolecular networks based on the classification method.

6.5 App implementation

6.5.1 CytoCtrlAnalyser architecture

CytoCtrlAnalyser is a Cytoscape app that is implemented based on the Open Service Gateway Initiative (OSGi) framework. OSGi is a Java framework for developing and deploying modular software programs and libraries. The recent versions of Cytoscape platform has adopted OSGi technology [183]. Therefore, both core modules and Apps in Cytoscape 3.x are OSGi bundles, which reduces the complexity of developing Apps remarkably. In addition, Cytoscape is a software developed in Java. Therefore, Cytoscape takes the advantage of Java that can be run in different operating systems with the Java virtual machine (JVM). The relationships among CytoCtrlAnalyser, Cytoscape and their running environments are shown in Fig. 6.9.

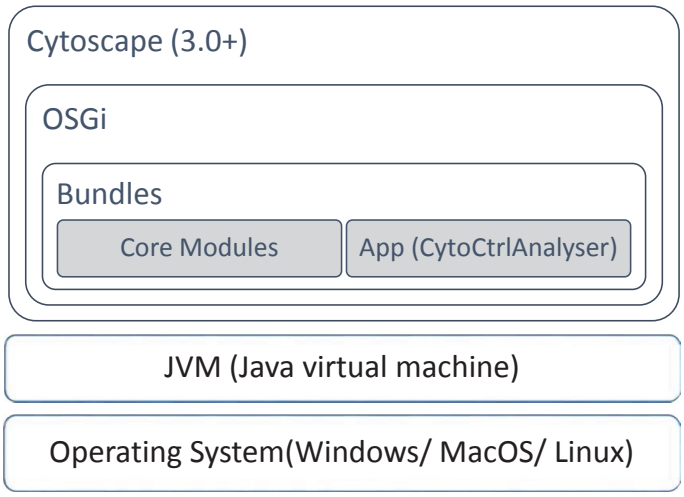


Figure 6.9: Relationships among the CytoCtrlAnalyser, Cytoscape and their running environments.

To achieve the functions of CytoCtrlAnalyser, there are three main functions modules in the CytoCtrlAnalyser, which are listed below.

1. Interface module: After opening CytoCtrlAnalyser, only one panel is created, which is the CytoCtrlAnalyser control panel. Users are able to check the checkbox to select the controllability algorithms they would like to run on the network loaded in Cytoscape. After pressing the *Analyze* button, the results are shown in the node table of table panel.

2. Task manage module: This module has two functions. The first is to receive the instructions from Cytoscape and open / close the CytoCtrlAnalyser control panel. Second, the module monitors action of the *Analyze* button on the control panel and calls algorithms accordingly.

3. Algorithm module: There are two parts in this module. The first part includes a group of algorithms related to the network control. Since the network control problems are formulated as graph-theoretic

problems, the second part includes several classical algorithms in graph theorem which are called by the controllability algorithms.

6.5.2 Relationships among the controllability algorithms

There are nine network controllability algorithms implemented in CytoCtrlAnalyser. The relationships among these algorithms can be represented by Fig. 6.10.

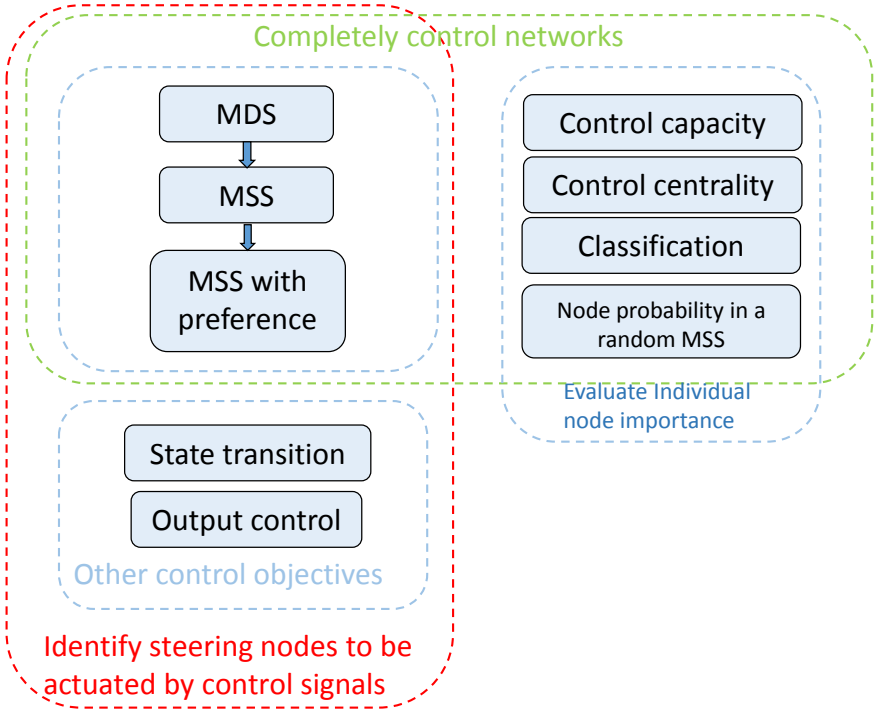


Figure 6.10: Relationships among network controllability algorithms.

From Fig. 6.10 we can see that the algorithms for MDS, MSS, MSS with preference, output controllability and state transittability are designed to identify a set of nodes which should be actuated by control signals such that different control objectives could be achieved. The algorithms for control capacity, control centrality, node classification and node probability in random MSS are designed to evaluate the importance of nodes from different aspects in the controllability of networks. Besides algorithms for output controllability and state transittability, all the algorithms are designed based on the completely controllability of networks. The arrows from MDS to MSS and from MSS to MSS with preference indicate the progress of the research on network controllability. Controlling an MDS is a necessary condition for completely controlling a network. To further investigate the controllability, the MSS has been proposed, which is a sufficient and necessary condition for completely controlling a network. Since MSSs of a network are not unique, the MSS with preference has been studied.

To investigated the controllability of networks, the studies have formulated the problems in control theory to the graph theoretic problems based on the structural control theorem. Therefore, nine network controllability algorithms are implemented based on classical graph-theoretic algorithms. The relationships among the network controllability algorithms and graph-theoretic algorithms are illustrated in Fig. 6.11.

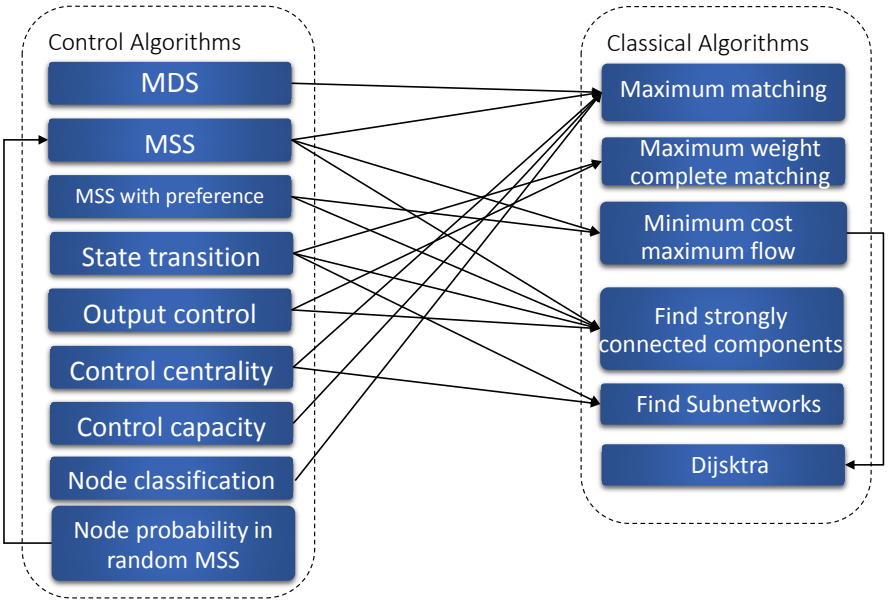


Figure 6.11: Calling relationship of algorithms.

6.5.3 Algorithm implementation

Since most algorithms are illustrated comprehensively in the related papers [19, 56, 62, 63, 67, 70, 74, 101, 103], this section only gives brief explanations for some specific problems during implementation.

When running the algorithms, CytoCtrlAnalyser acquires network information from Cytoscape and create a copy of the network, which is stored in the form of adjacent list. Therefore, all graph-theoretic algorithms are implemented based on the adjacent list in CytoCtrlAnalyser, which would enable the controllability algorithms to run on large networks and have higher efficiency for sparse networks.

MDS

The identification of MDS can be formulated as a maximum cardinality bipartite matching problem in a bipartite graph corresponding to the original network. The detail about the MDS and the identification algorithm can be referred in [19]. In CytoCtrlAnalyser, Hopcroft?CKarp algorithm [71] is employed to solve the maximum cardinality matching problem.

It should be mentioned that the MDSs of a network are not unique. For the algorithm of identifying MDS,

there are no stochastic variables when CytoCtrlAnalyser copies the network from Cytoscape and identifies the MDS. Therefore, for a given network imported to Cytoscape, the CytoCtrlAnalyser always returns a same MDS. In addition, it should be noticed that the identified MDSs of a network is related to the node order or edge order in the adjacent list created by CytoCtrlAnalyser. If a network is input with different orders of nodes or edges, CytoCtrlAnalyser may identifies different MDSs of the network.

MSS and MSS with preference

The identification of MSS has been formulated to a minimum cost maximum flow (MCMF) problem in a digraph which is constructed according to the topology of the network. The detail of the algorithm can be referred in [62] and the algorithm for solving the MCMF problem can be found in [73]. Notice that for each MSS, there is a subset of the MSS being an MDS. Then we firstly identify an MDS to improve the algorithm efficiency during the implementation of the algorithm. The identified MDS corresponds to a zero cost flow in the constructed digraph and the final MCMF could be augmented from the zero cost flow. Similar to the identification of MDS, there are no stochastic variables when CytoCtrlAnalyser retrieves the network from Cytoscape and identifies the MSS. Therefore, for a given network imported to Cytoscape, the CytoCtrlAnalyser always returns a same MSS. If different MSSs are needed, the network with different orders of nodes or edges should be input to CytoCtrlAnalyser.

The digraph constructed for the identification of MSS can also be applied to identify MSS with preference. However, to identify MSS with preference, costs of some edges in the digraph should be modified according to the preference values of nodes. The MCMF in the modified digraph indicates the MSS with preference. The detailed description of the algorithm can be referred in [92]. In CytoCtrlAnalyser, the users need to indicate the column which stores the preference values of nodes in the node table. It is possible that more than one MSSs in a network have the same maximum preference value. In this situation, CytoCtrlAnalyser always returns a same MSS with maximum preference. If different MSSs with maximum preference are needed, the network should be input to CytoCtrlAnalyser with different orders of nodes or edges.

Transittability and output controllability

Both the algorithms for transittability and output controllability require the users to indicate a column in node table with Boolean values. The nodes whose states are intended to be changed by the users are assigned true values. The difference between investigating the structural transittablilty and structural output controllability is that when steering the states of selected nodes, the structural transittablilty does not change the states of other nodes while output controllability does not consider the states of other nodes.

The identification of steering nodes for transittability and output controllability have been formulated to maximum weight complete matching problem in two different bipartite graphs constructed according to network topology and customized control settings. In CytoCtrlAnalyser, the maximum weight complete matching problem is solved by the Kuhn-Munkres (KM) algorithm [81]. Detail description of constructing

the weighted bipartite graph for transittability and output controllability can be referred in [56] and [70], respectively.

Control capacity and probabilities of nodes in a random MSS

Both MDSs and MSSs of a network are not unique. Therefore, to understand the roles of nodes played in controllability of networks, it is worthwhile studying the probabilities that the nodes would appear in a random MDS / MSS. Control capacity is to MDSs of networks as probability in an MSS is to MSSs.

The control capacity measures the likelihood of each node appearing in a random MDS. The algorithm to calculate the control capacity can be referred in [101]. The algorithm iterates many times to randomly sample MDSs of a network. Each iteration identifies one MDS and the MDSs identified by different iterations could be identical. Then the likelihood is the ratio of the times that a node appearing in MDSs to the times of iterations. In [101], the authors claimed that by iterating $T = NlnN$ times, where N is the number of nodes in the network, the sampling results converge to the actual values. However, when N is small, the number of iterations could not take enough samples of MDSs and the results do not converge to the actual values. Therefore, for the implementation of the algorithm, the value T is defined as $Max(NlnN, 1000)$, which is the larger value between $NlnN$ and 1000.

For the calculation of nodes' probabilities in a random MSS, there is no algorithm developed to sample MSSs uniformly at present. CytoCtrlAnalyser implements a method proposed in [92] to sample MSSs of networks. To sample different MSSs, the algorithm in CytoCtrlAnalyser intentionally exchanges the orders of nodes or edges when retrieving network topology from Cytoscape. The algorithm would return different MSSs since the orders of nodes or edges are different. Due to the efficiency and running time, the method samples 1000 MSSs at each execution.

Control centrality

Control centrality is developed to quantify the ability of a node to control a network, which equals to the dimension of the controllable subspace. In other words, by regulating only one node with control signal, the maximum number of nodes whose states can be steered from any initial state to any final state in the network is equal to the control centrality of the node. The CytoCtrlAnalyser integrates the algorithm to calculate the control centrality that is proposed in [63].

Node classification

The MDSs of a network are not unique, but the cardinality of the MDSs are the same. Based on the effect of removing a node to the cardinality of the MDSs, the paper [103] classified the nodes as indispensable, neutral or dispensable, which correlate to increasing, no effect, or decreasing the cardinality of the MDSs of the network by removing the node and edges that connect to the node.

6.6 Network dynamic model and structural controllability theorems

6.6.1 System dynamic model and graph representation

In CytoCtrlAnalyser, all implemented algorithms are based on networks with linear dynamics. Although dynamics of biological systems are nonlinear, the controllability of nonlinear systems is in many aspects structurally similar to that of linear systems. First, investigating controllability of locally linearized system is the first step to ultimately develop control strategies for complex nonlinear networks [40]. In addition, if a network is structurally controllable, then it is completely controllable for almost all possible parameter realizations [20]. Therefore, the structural controllability of linear system can provide a sufficient condition for controllability for most nonlinear systems [19, 51]. Recently, by applying structural linear controllability theorems to nonlinear C. elegans neuron network, researchers predicted the involvement of each C. elegans neuron in locomotor behaviors and then verified their prediction by experiments [94], which provided a directly experimental proof of the feasibility of developed structural controllability theorems.

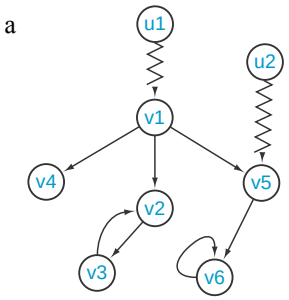
In this study, the dynamics of a network with n nodes is represented by the linear time-invariant dynamic model, which is described by the equation:

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \tag{6.1}$$

where $\mathbf{x}(t) = (x_1(t), ..., x_n(t))^T$ is a state vector that describes the states of all the nodes in the network. A is an $n \times n$ state transition matrix, which is determined by the adjacent matrix of the network, indicating the regulatory relationships between nodes in the network. $\mathbf{u}(t)$ is an input vector of m independent input control signals. The $n \times m$ matrix B is an input matrix that indicates the nodes which are directly actuated by input control signals. The network with the control signals described by the equation (6.1) is denoted as system (A, B).

Each system (A, B) has a corresponding graph representation G(A,B) (See Fig. 6.12). G(A,B) is a digraph which contains nodes $V_A = \{v_1, ..., v_n\}$ and $V_U = \{u_1, ..., u_m\}$ as well as edges $v_j \to v_i$ for $a_{ij} \neq 0$ and $u_j \to v_i$ for $b_{ij} \neq 0$. The original network without input control signals is denoted as G(A), which is a subgraph of G(A, B) induced by the node set V_A, where nodes in V_A correspond to the nodes in the original network. The edges between nodes in V_A are indicated by the state transition matrix A, which indicate the regulatory interactions between nodes. The nodes in V_U represent input nodes, which correspond to external control signals or environmental stimuli to the network. Each node u_i in V_U of G(A,B) corresponds to the entry $u_i(t)$ in $\mathbf{u}(t)$. Edges from a node u_i in V_U to nodes in V_A are indicated by the ith column in the input matrix B. Fig. 6.12 is an example of the system (A, B) and its graph representation.

For many real complex networks, such as biomolecular networks, it is feasible to qualify whether there is a regulatory relationship between two nodes (biomolecules), but it is difficult to quantify the strength of



b

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ a_{21} & 0 & a_{23} & 0 & 0 & 0 \\ 0 & a_{32} & 0 & 0 & 0 & 0 \\ a_{41} & 0 & 0 & 0 & 0 & 0 \\ a_{51} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{65} & a_{66} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & b_{52} \\ 0 & 0 \end{bmatrix}$$

Figure 6.12: Graph representation of system (A, B). (a): $G(A, B)$ corresponds to system (A, B). (b): The state transition matrix and input matrix of the system (A, B).

the regulation. Therefore, the concept of the structural system has been applied to study the dynamics of network systems [20]. System (A, B) is called a structural system when the entries in matrices (A, B) are either fixed zero or independent parameters. The following subsections give some introduction to the control problems based on structural systems.

6.6.2 Completely structural controllability

A network is completely controllable if it can be steered from any initial state $\mathbf{x}_0$ to any desired final state $\mathbf{x}_1$ in finite time with appropriate control signals. According to the Kalman's controllability rank condition, system (A, B) is completely controllable if and only if the $n \times nm$ controllability matrix

$$\mathfrak{C} = \begin{bmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{bmatrix} \tag{6.2}$$

has full row rank n [34]. For structural systems, we say the structural system (A, B) is completely structurally controllable if it is possible to choose the values for the independent entries in matrices A and B such that the Kalman's controllability rank condition is satisfied [20].

The graph-theoretic conditions for structural controllability have been developed in previous studies [20, 120]. Before introducing structural controllability theorem, we introduce two following concepts.

***Definition* 6.1** (Inaccessibility [20, 62])**.** A node v_i in the digraph G(A,B) is called inaccessible if and only if there exist no directed paths reaching v_i from an input vertex in V_U.

***Definition* 6.2** (Dilation [20, 62])**.** The digraph G(A,B) contains a dilation if and only if there is a subset S of V_A such that $|T(S)| < |S|$, where $T(S) = \{v_j \ | (v_j \to v_i) \in E(G),\ v_i \in S\}$ and E(G) is the edge set of G(A,B). $|S|$ and $|T(S)|$ are the cardinality of set S are T(S), respectively.

***Theorem* 6.1** (Structural controllability theorem [20, 62])**.** A structural system (A, B) is completely structurally controllable if and only if:

i) no dilation in digraph G(A,B).

ii) no inaccessible node in V_A.

6.6.3 Structural output controllability

The output of a linear dynamic system (A, B) can be described by the following equation:

$$\mathbf{y}(t) = C\mathbf{x}(t) \tag{6.3}$$

where $\mathbf{y}(t) = (y_1(t), ..., y_p(t))^T$ is an output vector of the network and each entry represents an output. The outputs of the network are linear combinations of node states in the network represented by the $p \times n$ matrix C. A system described by equations (6.1) and (6.3) is denoted by matrix triplet (A, B, C).

In this study, the outputs of a network is defined as states of a set of nodes. Based on this definition of output, there is one and only one nonzero entry in each row of C and the nonzero entry indicates one node in the network as an output. Then $\mathbf{y}(t)$ is a p-dimensional vector that each entry corresponds to state of one node. By defining the outputs in this way, the output controllability is basically the same as the concept of target controllability [68].

A system is output controllable if for any initial output vector $\mathbf{y}_0 = \mathbf{y}(t_0)$ and any other final output vector $\mathbf{y}_1$, there exists a finite time t_f and inputs $\mathbf{u}(t)$, such that $\mathbf{y}(t_f) = \mathbf{y}_1$. For a system (A, B, C), the $p \times mn$ output controllability matrix is defined as:

$$\mathfrak{o}\mathfrak{C} = \begin{bmatrix} CB & CAB & CA^2B & ... & CA^{n-1}B \end{bmatrix}. \tag{6.4}$$

Based on the control theory, system (A, B, C) is output controllable if and only if $\text{rank}\,(\mathfrak{o}\mathfrak{C}) = p$ [69].

For structural systems, by arbitrarily choosing the value of free parameters in A, B and C, the rank of $\mathfrak{o}\mathfrak{C}$ can reach a maximum value. The maximum value is defined as the *generic dimension of the controllable output subspace* of structural system (A, B, C) and denoted by $GDCOS(A, B, C)$ [24, 70]. The structural system (A, B, C) is structurally output controllable if $GDCOS(A, B, C) = p$.

6.6.4 Structural transittability

If there exists input control signals $\mathbf{u}(t),\ t \in [0,\ t_f]$, by which the system (A, B) can be transited between two specific states $\mathbf{x}_0$ and $\mathbf{x}_1$. The system (A, B) is called transittable between these two specific states.

For structural matrix M, if a matrix $\tilde{M}$ can be obtained by fixing the independent entries of M at some specific values, the matrix $\tilde{M}$ is called admissible with respect to M. Considering a structural system (A, B), the state vector $\mathbf{x}$ is a structural vector, in which entries are independent parameters or fixed zeros. A structural system (A, B) is called structural transittable between two structural states $\mathbf{x}_0$ and $\mathbf{x}_1$ if and only if there exist matrices $\tilde{A}$, $\tilde{B}$, $\tilde{\mathbf{x}}_0$ and $\tilde{\mathbf{x}}_1$ which are admissible with respect to A, B, $\mathbf{x}_0$ and $\mathbf{x}_1$, respectively, such that the system $(\tilde{A}, \tilde{B})$ is transittable between $\tilde{\mathbf{x}}_0$ and $\tilde{\mathbf{x}}_1$ [56].

For structural system (A, B), by arbitrarily choosing the value of free parameters in A and B, the rank of controllability matrix $\mathfrak{C}$ can reach a maximum value, which is denoted as *generic dimension of controllable subspace* $GDCS(A, B)$. Then we have the following theorem for structural transittability:

***Theorem* 6.2.** (Structural transittability theorem) [56] The structure system (A, B) is structurally transittable between two specific structural states $\mathbf{x}_0$ and $\mathbf{x}_1$ with either belonging to $span\{\mathfrak{C}\}$, if and only if

$$GDCS(A, \bar{B}) = GDCS(A, B),$$

where $\bar{B} = [\mathbf{x}_0 - \mathbf{x}_1, B]$.

7 Summary, contributions and future work

7.1 Summary

In order to control states of complex networks, the first challenge is to identify the minimum set of steering nodes which should be actuated by input control signals. For biomolecular networks, the difficulties come from the lack of accurate dynamic models and methods for analysing the controllability of dynamic models. Recent progress on the controllability of general complex networks have been made based on structural controllability theorems. To better understand the controllability of biomolecular networks, this thesis aims to identify the minimum set of steering nodes in different scenarios of controlling biomolecular networks. In addition, a software system which integrates controllability algorithms is implemented.

Chapter 2 provides a comprehensive review of network controllability and achieves Objective 1. It summarizes methods to investigate controllability of complex networks and the applications to biological networks. For different methods, the motivations and the application scenarios are discussed.

Chapter 3 studies the minimum number of steering nodes required to completely control complex networks. The minimum steering node set is denoted as MSS. By comparing the biological significance of the MSSs and the MDSs, which is a commonly used concept in network controllability studies, we conclude that MSSs are more critical in the dynamics of biomolecular networks. Objective 2 is accomplished by Chapter 3.

Chapter 4 and Chapter 5 fulfil Objective 3 and Objective 4, respectively. Objective 3 and Objective 4 are motivated by realistic demands or constraints on the steering nodes. On one hand, Objective 3 aims at improving the efficiency of control by using fewer steering nodes to achieve control objectives. Based on this idea, an algorithm is proposed to identify steering nodes for output controllability of complex networks in Chapter 4. Output controllability measures the ability of controlling the states of subsets of nodes in networks by actuating the steering nodes. Compared to completely controlling a whole network, controlling subsets of nodes in the network requires fewer steering nodes, which is more efficient. On the other hand, for controlling biomolecular networks, chemical molecules such as drugs are the most feasible types of input control signals. Since not all biomolecules have the same chance to be actuated by available drugs, Objective 4 intends to increase the likelihood that the identified steering nodes can be actuated by available drugs. Chapter 5 fulfills this objective by developing an algorithm to identify steering nodes with preference. When preference values are assigned to nodes according to their abilities to bind to drugs, the developed algorithm

can identify a steering node set which is most likely to be actuated by drugs as input control signals.

Chapter 6 introduces a software system called CytoCtrlAnalyser. CytoCtrlAnalyser is a Cytoscape app for analysing controllability of complex networks. Nine network controllability algorithms for: identification of (1) MDS [19], (2) MSS [62], (3) MSS with preference [74], (4) steering nodes for transittability [56] and (5) steering nodes for output controllability [67, 70], and calculation of (6) control centrality [182], (7) control capacity [101], (8) node classification [103] and (9) probability of each node in a random MSS [92], are integrated in the current version of CytoCtrlAnalyser.

7.2 Future work

Based on the work presented in the thesis, some directions of future work are proposed as follows:

1. Identifying steering nodes sets with drug binding preference for different control objectives.
   In the thesis, an algorithm was developed to identify an MSS based on certain pre-defined preferences. Since the MSS is one type of steering node set for complete controllability and there are other types of steering node sets for different control objectives such as output controllability and transittability, algorithms to identify steering node sets with preference for different control objectives are needed. For example, the output controllability can be applied to the drug target identification in biomolecular networks and the identified steering nodes are considered as potential drug targets. If the steering node sets are identified based on the drug binding preference, the potential drug targets are more likely to be actuated by available drugs.

2. Extending structural controllability theorems for better dynamic models of biomolecular networks.
   The thesis studies controllability of biomolecular networks based on structural controllability theorems. However, biomolecular networks show some properties that make directly using structural controllability theorems inappropriate. For example, structural controllability makes a general assumption that all the nonzero entries in state transition matrices of biomolecular networks are independent. However, regulatory interactions between biomolecules may depend on each other, which suggests that the nonzero entries in state transition matrices may not be independent. Thus algorithms should be developed to study the controllability of networks with dependent nonzero entries in state transition matrices.

3. Considering trajectories of control processes.
   Some states of the biomolecular networks may be lethal for the biomolecular systems. When steering biomolecular networks from certain states to other desired states, the trajectories of transition processes should avoid these forbidden states. Therefore, methods should be developed to determine control strategies to avoid some forbidden states in state transition trajectories and to find optimal steering node sets based on the strategies.

4. Investigating observability of biomolecular networks.

   Observability is a mathematical dual problem of controllability, which measures the ability of inferring states of nodes in a network by monitoring the states of some nodes in the network. Since there is much similarity between controllability and observability mathematically, the methods developed for network controllability can be extend to network observability. By observing states of a specific set of nodes, the states of other nodes or the whole network can be inferred, which suggests promising applications such as disease diagnosis and early disease detection.

References

[1] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress, "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes," *Human Molecular Genetics*, vol. 23, no. 22, pp. 5866–5878, 2014.

[2] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, p. 56, 2011.

[3] L.-Z. Liu, F.-X. Wu, and W.-J. Zhang, "Reverse engineering of gene regulatory networks from biological data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 5, pp. 365–385, 2012.

[4] A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, and E. E. Wanker, "A directed protein interaction network for investigating intracellular signal transduction," *Sciencs Signaling*, vol. 4, no. 189, pp. rs8–rs8, 2011.

[5] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences*, vol. 98, no. 8, pp. 4569–4574, 2001.

[6] P. Csermely, V. Agoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends in Pharmacological Sciences*, vol. 26, no. 4, pp. 178–182, 2005.

[7] Y.-F. Dai and X.-M. Zhao, "A survey on the computational approaches to identify drug targets in the postgenomic era," *BioMed Research International*, vol. 2015, 2015.

[8] X. Wang, N. Gulbahce, and H. Yu, "Network-based methods for human disease gene prediction," *Briefings in Functional Genomics*, vol. 10, no. 5, pp. 280–293, 2011.

[9] B. Chen, W. Fan, J. Liu, and F.-X. Wu, "Identifying protein complexes and functional modules - from static ppi networks to dynamic ppi networks," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 177–194, 2013.

[10] L. Chen, R.-S. Wang, and X.-S. Zhang, *Biomolecular networks: methods and applications in systems biology*. John Wiley & Sons, 2009, vol. 10.

[11] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang, "The yeast cell-cycle network is robustly designed," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 14, pp. 4781–4786, 2004.

[12] E. Klipp and W. Liebermeister, "Mathematical modeling of intracellular signaling pathways," *BMC Neuroscience*, vol. 7, no. 1, p. S10, 2006.

[13] M. I. Davidich and S. Bornholdt, "Boolean network model predicts cell cycle sequence of fission yeast," *PLoS One*, vol. 3, no. 2, p. e1672, 2008.

[14] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," *Molecular Systems Biology*, vol. 4, no. 1, p. 228, 2008.

[15] Z. Li, R.-S. Wang, X.-S. Zhang, and L. Chen, "Detecting drug targets with minimum side effects in metabolic networks," *IET Systems Biology*, vol. 3, no. 6, pp. 523–533, 2009.

[16] Z. Li, R.-S. Wang, and X.-S. Zhang, "Two-stage flux balance analysis of metabolic networks for drug target identification," *BMC Systems Biology*, vol. 5, no. 1, p. S11, 2011.

[17] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, no. 5, pp. 60–69, 2003.

[18] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, p. 440, 1998.

[19] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Controllability of complex networks," *Nature*, vol. 473, no. 7346, p. 167, 2011.

[20] C.-T. Lin, "Structural controllability," *IEEE Transactions on Automatic Control*, vol. 19, no. 3, pp. 201–208, 1974.

[21] A. Gibbons, *Algorithmic graph theory*. Cambridge university press, 1985.

[22] S. Hosoe, "Determination of generic dimensions of controllable subspaces and its application," *IEEE Transactions on Automatic Control*, vol. 25, no. 6, pp. 1192–1196, 1980.

[23] S. Poljak, "On the generic dimension of controllable subspaces," *IEEE Transactions on Automatic Control*, vol. 35, no. 3, pp. 367–369, 1990.

[24] K. Murota and S. Poljak, "Note on a graph-theoretic criterion for structural output controllability," *IEEE Transactions on Automatic Control*, vol. 35, no. 8, pp. 939–942, 1990.

[25] N. J. Cowan, E. J. Chastain, D. A. Vilhena, J. S. Freudenberg, and C. T. Bergstrom, "Nodal dynamics, not degree distributions, determine the structural controllability of complex networks," *PLoS One*, vol. 7, no. 6, p. e38398, 2012.

[26] Z. Yuan, C. Zhao, Z. Di, W.-X. Wang, and Y.-C. Lai, "Exact controllability of complex networks," *Nature Communications*, vol. 4, p. 2447, 2013.

[27] A. Cho, "Scientific link-up yields 'control panel' for networks," *Science*, vol. 332, no. 6031, pp. 777–777, 2011.

[28] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[29] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction," *Journal of Molecular Biology*, vol. 311, no. 4, pp. 681–692, 2001.

[30] M. S. Dasika, A. Gupta, and C. D. Maranas, "A mixed integer linear programming (milp) framework for inferring time delay in gene regulatory networks," in *Biocomputing 2004*. World Scientific, 2003, pp. 474–485.

[31] J. Zhao, Y. Zhou, X. Zhang, and L. Chen, "Part mutual information for quantifying direct associations in networks," *Proceedings of the National Academy of Sciences*, vol. 113, no. 18, pp. 5130–5135, 2016.

[32] S. Gu, F. Pasqualetti, M. Cieslak, Q. K. Telesford, B. Y. Alfred, A. E. Kahn, J. D. Medaglia, J. M. Vettel, M. B. Miller, S. T. Grafton *et al.*, "Controllability of structural brain networks," *Nature Communications*, vol. 6, 2015.

[33] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[34] R. E. Kalman, "Mathematical description of linear dynamical systems," *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, vol. 1, no. 2, pp. 152–192, 1963.

[35] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[36] E. R. Alvarez-Buylla, M. Benítez, E. B. Dávila, A. Chaos, C. Espinosa-Soto, and P. Padilla-Longoria, "Gene regulatory network models for plant development," *Current Opinion in Plant Biology*, vol. 10, no. 1, pp. 83–91, 2007.

[37] J. Krumsiek, C. Marr, T. Schroeder, and F. J. Theis, "Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network," *PLoS One*, vol. 6, no. 8, p. e22649, 2011.

[38] D. Cheng and H. Qi, "Controllability and observability of boolean control networks," *Automatica*, vol. 45, no. 7, pp. 1659–1667, 2009.

[39] E. Fornasini and M. E. Valcher, "Observability, reconstructibility and state observers of boolean control networks," *IEEE Transactions on Automatic Control*, vol. 58, no. 6, pp. 1390–1401, 2013.

[40] Y.-Y. Liu and A.-L. Barabási, "Control principles of complex systems," *Reviews of Modern Physics*, vol. 88, no. 3, p. 035006, 2016.

[41] T. Akutsu, M. Hayashida, W.-K. Ching, and M. K. Ng, "Control of boolean networks: Hardness results and algorithms for tree structured networks," *Journal of Theoretical Biology*, vol. 244, no. 4, pp. 670–679, 2007.

[42] J. Kim, S.-M. Park, and K.-H. Cho, "Discovery of a kernel for controlling biomolecular regulatory networks," *Scientific Reports*, vol. 3, p. 2223, 2013.

[43] C.-Y. Huang and J. E. Ferrell, "Ultrasensitivity in the mitogen-activated protein kinase cascade," *Proceedings of the National Academy of Sciences*, vol. 93, no. 19, pp. 10 078–10 083, 1996.

[44] R. Khanin, V. Vinciotti, and E. Wit, "Reconstructing repressor protein levels from expression of gene targets in escherichia coli," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18 592–18 596, 2006.

[45] G. Haynes and H. Hermes, "Nonlinear controllability via lie theory," *SIAM Journal on Control*, vol. 8, no. 4, pp. 450–460, 1970.

[46] H. J. Sussmann and V. Jurdjevic, "Controllability of nonlinear systems," *Journal of Differential Equations*, vol. 12, no. 1, pp. 95–116, 1972.

[47] R. Hermann and A. Krener, "Nonlinear controllability and observability," *IEEE Transactions on Automatic Control*, vol. 22, no. 5, pp. 728–740, 1977.

[48] A. Isidori, *Nonlinear control systems*. Springer Science & Business Media, 2013.

[49] S. P. Cornelius, W. L. Kath, and A. E. Motter, "Realistic control of network dynamics," *Nature Communications*, vol. 4, p. 1942, 2013.

[50] P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi *et al.*, "Linear modeling of mrna expression levels during cns development and injury." in *Pacific Symposium on Biocomputing*, vol. 4, no. 1, 1999, pp. 41–52.

[51] J.-J. E. Slotine, W. Li *et al.*, *Applied nonlinear control*. prentice-Hall Englewood Cliffs, NJ, 1991, vol. 199, no. 1.

[52] R. Shields and J. Pearson, "Structural controllability of multiinput linear systems," *IEEE Transactions on Automatic control*, vol. 21, no. 2, pp. 203–212, 1976.

[53] K. Glover and L. Silverman, "Characterization of structural controllability," *IEEE Transactions on Automatic control*, vol. 21, no. 4, pp. 534–537, 1976.

[54] S. Hosoe and K. Matsumoto, "On the irreducibility condition in the structural controllability theorem," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 963–966, 1979.

[55] A. Linnemann, "A further simplification in the proof of the structural controllability theorem," *IEEE Transactions on Automatic Control*, vol. 31, no. 7, pp. 638–639, 1986.

[56] F.-X. Wu, L. Wu, J. Wang, J. Liu, and L. Chen, "Transittability of complex networks and its applications to regulatory biomolecular networks," *Scientific Reports*, vol. 4, p. 4819, 2014.

[57] T. Nepusz and T. Vicsek, "Controlling edge dynamics in complex networks," *Nature Physics*, vol. 8, no. 7, p. 568, 2012.

[58] N. J. Cowan, E. J. Chastain, D. A. Vilhena, J. S. Freudenberg, and C. T. Bergstrom, "Nodal dynamics, not degree distributions, determine the structural controllability of complex networks," *PLoS One*, vol. 7, no. 6, p. e38398, 2012.

[59] S. Nie, X. Wang, H. Zhang, Q. Li, and B. Wang, "Robustness of controllability for networks based on edge-attack," *PLoS One*, vol. 9, no. 2, p. e89066, 2014.

[60] W.-X. Wang, X. Ni, Y.-C. Lai, and C. Grebogi, "Optimizing controllability of complex networks by minimum structural perturbations," *Physical Review E*, vol. 85, no. 2, p. 026115, 2012.

[61] L. Wu, M. Li, J. Wang, and F.-X. Wu, "Cytoctrlanalyser: a cytoscape app for biomolecular network controllability analysis," *Bioinformatics*, vol. 34, no. 8, pp. 1428-1430 2017.

[62] L. Wu, M. Li, J. Wang, and F.-X. Wu, "Minimum steering node set of complex networks and its applications to biomolecular networks," *IET Systems Biology*, vol. 10, no. 3, pp. 116–123, 2016.

[63] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Control centrality and hierarchical structure in complex networks," *PLoS One*, vol. 7, no. 9, p. e44459, 2012.

[64] F. L. Iudice, F. Garofalo, and F. Sorrentino, "Structural permeability of complex networks to control signals," *Nature Communications*, vol. 6, 2015.

[65] X. Liu and L. Pan, "Controllability of the better chosen partial networks," *Physica A: Statistical Mechanics and its Applications*, vol. 456, pp. 120–127, 2016.

[66] C. Commault, J. Van Der Woude, and T. Boukhobza, "On the fixed controllable subspace in linear structured systems," *Systems & Control Letters*, vol. 102, pp. 42–47, 2017.

[67] L. Wu, Y. Shen, M. Li, and F.-X. Wu, "Drug target identification based on structural output controllability of complex networks," in *Bioinformatics Research and Applications*. Springer, 2014, pp. 188–199.

[68] J. Gao, Y.-Y. Liu, R. M. D'Souza, and A.-L. Barabási, "Target control of complex networks," *Nature Communications*, vol. 5, 2014.

[69] K. Ogata, *Modern Control Engineering*, 3rd ed. Prentice-Hall, 1997.

[70] L. Wu, Y. Shen, M. Li, and F.-X. Wu, "Network output controllability-based method for drug target identification," *IEEE Transactions on Nanobioscience*, vol. 14, no. 2, pp. 184–191, 2015.

[71] J. E. Hopcroft and R. M. Karp, "An n^5/2 algorithm for maximum matchings in bipartite graphs," *SIAM Journal on Computing*, vol. 2, no. 4, pp. 225–231, 1973.

[72] X. Zhang, T. Lv, X. Yang, and B. Zhang, "Structural controllability of complex networks based on preferential matching," *PLoS One*, vol. 9, no. 11, p. e112039, 2014.

[73] M. T. Goodrich and R. Tamassia, *Algorithm design: foundation, analysis and internet examples*. John Wiley & Sons, 2006.

[74] L. Wu, L. Tang, M. Li, J. Wang, and F.-X. Wu, "The mss of complex networks with centrality based preference and its application to biomolecular networks," pp. 229–234, 2016.

[75] S. Pequito, S. Kar, and A. P. Aguiar, "On the complexity of the constrained input selection problem for structural linear systems," *Automatica*, vol. 62, pp. 193–199, 2015.

[76] G. Lindmark and C. Altafini, "Controllability of complex networks with unilateral inputs," *Scientific Reports*, vol. 7, 2017.

[77] W. J. Rugh and W. J. Rugh, *Linear system theory*. prentice hall Upper Saddle River, NJ, 1996, vol. 2.

[78] L.-Z. Wang, Y.-Z. Chen, W.-X. Wang, and Y.-C. Lai, "Physical controllability of complex networks," *Scientific Reports*, vol. 7, 2017.

[79] G. Li, P. Tang, C. Wen, and Z. Meng, "Boundary constraints for minimum cost control of directed networks," *IEEE Transactions on Cybernetics*, 2017.

[80] E. Czeizler, C. Gratie, W. K. Chiu, K. Kanhaiya, and I. Petre, "Target controllability of linear networks," in *International Conference on Computational Methods in Systems Biology*. Springer, 2016, pp. 67–81.

[81] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[82] X. Zhang, H. Wang, and T. Lv, "Efficient target control of complex networks based on preferential matching," *PLoS One*, vol. 12, no. 4, p. e0175375, 2017.

[83] X. Liu, L. Pan, H. E. Stanley, and J. Gao, "Controllability of giant connected components in a directed network," *Physical Review E*, vol. 95, no. 4, p. 042318, 2017.

[84] X. Piao, T. Lv, X. Zhang, and H. Ma, "Strategy for community control of complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 421, pp. 98–108, 2015.

[85] W.-F. Guo, S.-W. Zhang, Z.-G. Wei, T. Zeng, F. Liu, J. Zhang, F.-X. Wu, and L. Chen, "Constrained target controllability of complex networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2017, no. 6, p. 063402, 2017.

[86] R. Khazanchi, K. Dempsey, I. Thapa, and H. Ali, "On identifying and analyzing significant nodes in protein-protein interaction networks," in *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on.* IEEE, 2013, pp. 343–348.

[87] R. Badhwar and G. Bagler, "Control of neuronal network in caenorhabditis elegans," *PLoS One*, vol. 10, no. 9, p. e0139204, 2015.

[88] H. R. Noori, J. Schöttler, M. Ercsey-Ravasz, A. Cosa-Linan, M. Varga, Z. Toroczkai, and R. Spanagel, "A multiscale cerebral neurochemical connectome of the rat brain," *PLoS Biology*, vol. 15, no. 7, p. e2002612, 2017.

[89] K. Deisseroth, "Circuit dynamics of adaptive and maladaptive behaviour," *Nature*, vol. 505, no. 7483, pp. 309–317, 2014.

[90] M. L. Kringelbach, N. Jenkinson, S. L. Owen, and T. Z. Aziz, "Translational principles of deep brain stimulation," *Nature Reviews Neuroscience*, vol. 8, no. 8, pp. 623–635, 2007.

[91] M. Moes, A. Le Béchec, I. Crespo, C. Laurini, A. Halavatyi, G. Vetter, A. Del Sol, and E. Friederich, "A novel network integrating a mirna-203/snai1 feedback loop which regulates epithelial to mesenchymal transition," *PLoS One*, vol. 7, no. 4, p. e35440, 2012.

[92] L. Wu, L. Tang, M. Li, J. Wang, and F.-X. Wu, "Biomolecular network controllability with drug binding information," *IEEE Transactions on Nanobioscience*, vol. 16, no. 5, pp. 326–332, 2017.

[93] K. Kanhaiya, E. Czeizler, C. Gratie, and I. Petre, "Controlling directed protein interaction networks in cancer," *Scientific Reports*, vol. 7, no. 1, p. 10327, 2017.

[94] G. Yan, P. E. Vértes, E. K. Towlson, Y. L. Chew, D. S. Walker, W. R. Schafer, and A.-L. Barabási, "Network control principles predict neuron function in the caenorhabditis elegans connectome." *Nature*, 2017.

[95] L. Mendoza, "A network model for the control of the differentiation process in th cells," *Biosystems*, vol. 84, no. 2, pp. 101–114, 2006.

[96] H. J. Lee, N. Takemoto, H. Kurata, Y. Kamogawa, S. Miyatake, A. O'garra, and N. Arai, "Gata-3 induces t helper cell type 2 (th2) cytokine expression and chromatin remodeling in committed th1 cells," *Journal of Experimental Medicine*, vol. 192, no. 1, pp. 105–116, 2000.

[97] S. J. Szabo, S. T. Kim, G. L. Costa, X. Zhang, C. G. Fathman, and L. H. Glimcher, "A novel transcription factor, t-bet, directs th1 lineage commitment," *Cell*, vol. 100, no. 6, pp. 655–669, 2000.

[98] E. S. Hwang, S. J. Szabo, P. L. Schwartzberg, and L. H. Glimcher, "T helper cell fate specified by kinase-mediated interaction of t-bet with gata-3," *Science*, vol. 307, no. 5708, pp. 430–433, 2005.

[99] T. Jia, Y.-Y. Liu, E. Csóka, M. Pósfai, J.-J. Slotine, and A.-L. Barabási, "Emergence of bimodality in controlling complex networks," *Nature Communications*, vol. 4, p. 2002, 2013.

[100] X. Liu and L. Pan, "Identifying driver nodes in the human signaling network using structural controllability analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 2, pp. 467–472, 2015.

[101] T. Jia and A.-L. Barabási, "Control capacity and a random sampling method in exploring controllability of complex networks," *Scientific Reports*, vol. 3, p. 2354, 2013.

[102] X. Liu and L. Pan, "Detection of driver metabolites in the human liver metabolic network using structural controllability analysis," *BMC Systems Biology*, vol. 8, no. 1, p. 51, 2014.

[103] A. Vinayagam, T. E. Gibson, H.-J. Lee, B. Yilmazel, C. Roesel, Y. Hu, Y. Kwon, A. Sharma, Y.-Y. Liu, N. Perrimon *et al.*, "Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets," *PNAS*, vol. 113, no. 18, pp. 4976–4981, 2016.

[104] Y. Matsuoka, H. Matsumae, M. Katoh, A. J. Eisfeld, G. Neumann, T. Hase, S. Ghosh, J. E. Shoemaker, T. J. Lopes, T. Watanabe *et al.*, "A comprehensive map of the influenza a virus replication cycle," *BMC Systems Biology*, vol. 7, no. 1, p. 97, 2013.

[105] M. Uhart, G. Flores, and D. M. Bustos, "Controllability of protein-protein interaction phosphorylation-based networks: Participation of the hub 14-3-3 protein family," *Scientific Reports*, vol. 6, 2016.

[106] V. Ravindran, V. Sunitha, and G. Bagler, "Identification of critical regulatory genes in cancer signaling network using controllability analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 474, pp. 134–143, 2017.

[107] J. Ruths and D. Ruths, "Control profiles of complex networks," *Science*, vol. 343, no. 6177, pp. 1373–1376, 2014.

[108] B. Wang, L. Gao, Q. Zhang, A. Li, Y. Deng, and X. Guo, "Diversified control paths: A significant way disease genes perturb the human regulatory network," *PLoS One*, vol. 10, no. 8, p. e0135491, 2015.

[109] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Computational Biology*, vol. 6, no. 1, p. e1000641, 2010.

[110] B. Wang, L. Gao, and Y. Gao, "Control range: a controllability-based index for node significance in directed networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 04, p. P04011, 2012.

[111] B. Wang, L. Gao, Y. Gao, Y. Deng, and Y. Wang, "Controllability and observability analysis for vertex domination centrality in directed networks," *Scientific Reports*, vol. 4, 2014.

[112] M. E. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[113] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, p. 47, 2002.

[114] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4, pp. 175–308, 2006.

[115] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," *Molecular Systems Biology*, vol. 4, no. 1, p. 228, 2008.

[116] Y. Liu, J. Lu, and B. Wu, "Some necessary and sufficient conditions for the output controllability of temporal boolean control networks," *ESAIM. Control, Optimisation and Calculus of Variations*, vol. 20, no. 1, p. 158, 2014.

[117] Y. Liu, H. Chen, B. Wu, and L. Sun, "A mayer-type optimal control for multivalued logic control networks with undesirable states," *Applied Mathematical Modelling*, vol. 39, no. 12, pp. 3357–3365, 2015.

[118] Y. Liu, H. Chen, J. Lu, and B. Wu, "Controllability of probabilistic boolean control networks based on transition probability matrices," *Automatica*, vol. 52, pp. 340–345, 2015.

[119] H. H. Rosenbrock, "State-space and multivariable theory," 1970.

[120] R. Shields and J. Pearson, "Structural controllability of multiinput linear systems," *IEEE Transactions on Automatic control*, vol. 21, no. 2, pp. 203–212, 1976.

[121] A. Olshevsky, "Minimal controllability problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 3, pp. 249–258, Sept 2014.

[122] A. Olshevsky, "Minimum input selection for structural controllability," *arXiv preprint arXiv:1407.2884*, 2014.

[123] H. Yin and S. Zhang, "Minimum structural controllability problems of complex networks," *Physica A: Statistical Mechanics and its Applications*, 2015.

[124] J. Edmonds and R. M. Karp, "Theoretical improvements in algorithmic efficiency for network flow problems," *Journal of the ACM (JACM)*, vol. 19, no. 2, pp. 248–264, 1972.

[125] M. Sharir, "A strong-connectivity algorithm and its applications in data flow analysis," *Computers & Mathematics with Applications*, vol. 7, no. 1, pp. 67–72, 1981.

[126] P. Zhang, J. Iwasaki-Arai, H. Iwasaki, M. L. Fenyus, T. Dayaram, B. M. Owens, H. Shigematsu, E. Levantini, C. S. Huettner, J. A. Lekstrom-Himes *et al.*, "Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor c/ebpα," *Immunity*, vol. 21, no. 6, pp. 853–863, 2004.

[127] A. D. Friedman, "C/ebpα induces pu. 1 and interacts with ap-1 and nf-κb to regulate myeloid development," *Blood Cells, Molecules, and Diseases*, vol. 39, no. 3, pp. 340–343, 2007.

[128] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, p. 56, 2011.

[129] L. Chen, R. Wang, and X. Zhang, *Biomolecular Networks : Methods and Applications in Systems Biology*. John Wiley & Sons, 2009.

[130] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang, "Attack robustness and centrality of complex networks," *PLoS One*, vol. 8, no. 4, p. e59613, 2013.

[131] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, p. 651, 2000.

[132] E. Estrada, "Virtual identification of essential proteins within the protein interaction network of yeast," *Proteomics*, vol. 6, no. 1, pp. 35–40, 2006.

[133] A. Samal, S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain, "Low degree metabolites explain essential reactions and enhance modularity in biological networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 118, 2006.

[134] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Systems Biology*, vol. 6, no. 1, p. 87, 2012.

[135] J. Wang, W. Peng, and F.-X. Wu, "Computational approaches to predicting essential proteins: a survey," *Proteomics - Clinical Applications*, vol. 7, no. 1-2, pp. 181–192, 2013.

[136] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761supp, p. C47, 1999.

[137] B. Chen and F.-X. Wu, "Identifying protein complexes based on multiple topological structures in ppi networks," *IEEE Transactions on Nanobioscience*, vol. 12, no. 3, pp. 165–172, 2013.

[138] S. Zhang, X.-S. Zhang, and L. Chen, "Biomolecular network querying: a promising approach in systems biology," *BMC Systems Biology*, vol. 2, no. 1, p. 5, 2008.

[139] M. A. Lindsay, "Target discovery," *Nature Reviews Drug Discovery*, vol. 2, no. 10, p. 831, 2003.

[140] P. Csermely, V. Agoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends in Pharmacological Sciences*, vol. 26, no. 4, pp. 178–182, 2005.

[141] A. S. Azmi, Z. Wang, P. A. Philip, R. M. Mohammad, and F. H. Sarkar, "Proof of concept: network and systems biology approaches aid in the discovery of potent anticancer drug combinations," *Molecular Cancer Therapeutics*, vol. 9, no. 12, pp. 3137–3144, 2010.

[142] M. Kotlyar, K. Fortney, and I. Jurisica, "Network-based characterization of drug-regulated genes, drug targets, and toxicity," *Methods*, vol. 57, no. 4, pp. 499–507, 2012.

[143] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein–protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.

[144] E. van den Akker, B. Verbruggen, B. Heijmans, M. Beekman, J. Kok, E. Slagboom, and M. Reinders, "Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis," *Journal of Integrative Bioinformatics*, vol. 8, no. 2, pp. 222–238, 2011.

[145] B. Chen, J. Wang, M. Li, and F.-X. Wu, "Identifying disease genes by integrating multiple data sources," *BMC Medical Genomics*, vol. 7, no. 2, p. S2, 2014.

[146] B. Chen, M. Li, J. Wang, and F.-X. Wu, "Disease gene identification by using graph kernels and markov random fields," *Science China Life Sciences*, vol. 57, no. 11, pp. 1054–1063, 2014.

[147] W.-C. Hwang, A. Zhang, and M. Ramanathan, "Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery," *Clinical Pharmacology & Therapeutics*, vol. 84, no. 5, pp. 563–572, 2008.

[148] Z. Wu, X.-M. Zhao, and L. Chen, "A systems biology approach to identify effective cocktail drugs," vol. 4, no. 2, p. S7, 2010.

[149] J.-J. E. Slotine, W. Li *et al.*, *Applied nonlinear control*. prentice-Hall Englewood Cliffs, NJ, 1991.

[150] N. S.Nise, *Control System Engineering*, 6th ed. John Wiley & Sons, 2011.

[151] D. Jungnickel, *Graphs, Networks and Algorithms*, 3rd ed. Springer, 2005.

[152] M. Murakami and I. Kudo, "Prostaglandin e synthase: a novel drug target for inflammation and cancer," *Current Pharmaceutical Design*, vol. 12, no. 8, pp. 943–954, 2006.

[153] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[154] P. Sridhar, B. Song, T. Kahveci, and S. Ranka, *Mining metabolic networks for optimal drug targets*. World Scientific, 2008, pp. 291–302.

[155] X.-F. Zhang, L. Ou-Yang, Y. Zhu, M.-Y. Wu, and D.-Q. Dai, "Determining minimum set of driver nodes in protein-protein interaction networks," *BMC Bioinformatics*, vol. 16, no. 1, p. 146, 2015.

[156] X.-P. Zhang, F. Liu, and W. Wang, "Two-phase dynamics of p53 in the dna damage response," *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 8990–8995, 2011.

[157] T. Helikar, J. Konvalina, J. Heidel, and J. A. Rogers, "Emergent decision-making in biological signal transduction networks," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 1913–1918, 2008.

[158] J. M. Lee, E. P. Gianchandani, and J. A. Papin, "Flux balance analysis in the era of metabolomics," *Briefings in Bioinformatics*, vol. 7, no. 2, pp. 140–150, 2006.

[159] T. Jia, Y.-Y. Liu, E. Csóka, M. Pósfai, J.-J. Slotine, and A.-L. Barabási, "Emergence of bimodality in controlling complex networks," *Nature Communications*, vol. 4, p. 2002, 2013.

[160] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "Drugbank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D901–D906, 2008.

[161] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "Stitch: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D684–D688, 2008.

[162] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, "Pubchem substance and compound databases," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[163] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D115–D119, 2004.

[164] A. L. Hopkins and C. R. Groom, "The druggable genome," *Nature reviews Drug discovery*, vol. 1, no. 9, pp. 727–730, 2002.

[165] K. Lundstrom, "An overview on gpcrs and drug discovery: structure-based drug design and structural biology on gpcrs," *G Protein-Coupled Receptors in Drug Discovery*, pp. 51–66, 2009.

[166] G. Song, G. Ouyang, and S. Bao, "The activation of akt/pkb signaling pathway and cell survival," *Journal of Cellular and Molecular Medicine*, vol. 9, no. 1, pp. 59–71, 2005.

[167] S. Ramaswamy, N. Nakamura, F. Vazquez, D. B. Batt, S. Perera, T. M. Roberts, and W. R. Sellers, "Regulation of g1 progression by the pten tumor suppressor protein is linked to inhibition of the phosphatidylinositol 3-kinase/akt pathway," *Proceedings of the National Academy of Sciences*, vol. 96, no. 5, pp. 2110–2115, 1999.

[168] J. Chen, P. R. Somanath, O. Razorenova, W. S. Chen, N. Hay, P. Bornstein, and T. V. Byzova, "Akt1 regulates pathological angiogenesis, vascular maturation and permeability in vivo," *Nature Medicine*, vol. 11, no. 11, pp. 1188–1196, 2005.

[169] I. M. Ghobrial, A. Roccaro, F. Hong, E. Weller, N. Rubin, R. Leduc, M. Rourke, S. Chuma, A. Sacco, X. Jia *et al.*, "Clinical and translational studies of a phase ii trial of the novel oral akt inhibitor perifosine in relapsed or relapsed/refractory waldenström's macroglobulinemia," *Clinical Cancer Research*, vol. 16, no. 3, pp. 1033–1041, 2010.

[170] Y. Rao, R. Li, and D. Zhang, "A drug from poison: how the therapeutic effect of arsenic trioxide on acute promyelocytic leukemia was discovered," *Science China Life Sciences*, vol. 56, no. 6, pp. 495–502, 2013.

[171] F. Balkwill and A. Mantovani, "Inflammation and cancer: back to virchow?" *The Lancet*, vol. 357, no. 9255, pp. 539–545, 2001.

[172] J. Lu, H. Zeng, Z. Liang, L. Chen, L. Zhang, H. Zhang, H. Liu, H. Jiang, B. Shen, M. Huang *et al.*, "Network modelling reveals the mechanism underlying colitis-associated colon cancer and identifies novel combinatorial anti-cancer targets," *Scientific Reports*, vol. 5, 2015.

[173] J.-J. Lu, W. Pan, Y.-J. Hu, and Y.-T. Wang, "Multi-target drugs: the trend of drug research and development," *PLoS One*, vol. 7, no. 6, p. e40262, 2012.

[174] S. D. Markowitz and M. M. Bertagnolli, "Molecular basis of colorectal cancer," *New England Journal of Medicine*, vol. 361, no. 25, pp. 2449–2460, 2009.

[175] J. Rossi, S. Negrier, N. D. James, I. Kocak, R. Hawkins, H. Davis, U. Prabhakar, X. Qin, P. Mulders, and B. Berns, "A phase i/ii study of siltuximab (cnto 328), an anti-interleukin-6 monoclonal antibody, in metastatic renal cell cancer," *British Journal of Cancer*, vol. 103, no. 8, pp. 1154–1162, 2010.

[176] J. Karkera, H. Steiner, W. Li, V. Skradski, P. L. Moser, S. Riethdorf, M. Reddy, T. Puchalski, K. Safer, U. Prabhakar *et al.*, "The anti-interleukin-6 antibody siltuximab down-regulates genes implicated in tumorigenesis in prostate cancer patients from a phase i study," *The Prostate*, vol. 71, no. 13, pp. 1455–1465, 2011.

[177] R. D. Loberg, L. L. Day, J. Harwood, C. Ying, L. N. S. John, R. Giles, C. K. Neeley, and K. J. Pienta, "Ccl2 is a potent regulator of prostate cancer cell migration and proliferation," *Neoplasia*, vol. 8, no. 7, pp. 578–586, 2006.

[178] G. Soria and A. Ben-Baruch, "The inflammatory chemokines ccl2 and ccl5 in breast cancer," *Cancer Letters*, vol. 267, no. 2, pp. 271–285, 2008.

[179] S. Nagata, "Apoptosis by death factor," *Cell*, vol. 88, no. 3, pp. 355–365, 1997.

[180] R. M. Kluck, E. Bossy-Wetzel, D. R. Green, and D. D. Newmeyer, "The release of cytochrome c from mitochondria: a primary site for bcl-2 regulation of apoptosis," *Science*, vol. 275, no. 5303, pp. 1132–1136, 1997.

[181] X. Liu, Y. Wang, H. Ji, K. Aihara, and L. Chen, "Personalized characterization of diseases using sample-specific networks," *Nucleic Acids Research*, vol. 44, no. 22, pp. e164–e164, 2016.

[182] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Control centrality and hierarchical structure in complex networks," *PLoS One*, vol. 7, no. 9, p. e44459, 2012.

[183] J. Wang, J. Zhong, G. Chen, M. Li, F.-X. Wu, and Y. Pan, "Clusterviz: A cytoscape app for cluster analysis of biological network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 815–822, July 2015.

Appendix A

List of Publications

Refereed Journal Papers:

[1] **L. Wu**, M. Li, J. Wang, and F.-X Wu,"Cytoctrlanalyser: a cytoscape app for biomolecular network controllability analysis," *Bioinformatics*, vol. 34, no. 8, pp. 1428-1430, 2017.

[2] **L. Wu**, L. Tang, M. Li, J. Wang, and F.-X Wu, "Biomolecular network controllability with drug binding information," *IEEE Transactions on NanoBioscience*, vol. 16, no. 5, pp. 326-332, 2017.

[3] **L. Wu**, M. Li, J. Wang, and F.-X. Wu, "Minimum steering node set of complex networks and its applications to biomolecular networks," *IET Systems biology*, vol. 10, no. 3, pp. 116-123, 2016.

[4] **L. Wu**, Y. Shen, M. Li, and F.-X. Wu, "Network output controllability-based method for drug target identification," *IEEE Transactions on NanoBioscience*, vol. 14, no. 2, pp. 184-191, 2015.

[5] F.-X. Wu, **L. Wu**, J. Wang, J. Liu, and L. Chen, "Transittability of complex networks and its applications to regulatory biomolecular networks," *Scientific Reports*, vol. 4, 2014.

[6] L.-X. Li, **L. Wu**, H.-S. Zhang, and F.-X. Wu, "A fast algorithm for nonnegative matrix factorization and its convergence," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 10, pp. 1855-1863, 2014.

Refereed Conference Papers:

[1] **L. Wu**, L. Tang, M. Li, J. Wang, and F.-X. Wu, "The MSS of complex networks with centrality based preference and its application to biomolecular networks," *IEEE BIBM 2016*, pp. 229-234.

[2] **L. Wu**, Y. Shen, M. Li, and F.-X. Wu, "Drug target identification based on structural output controllability of complex networks," *ISBRA 2014*, pp. 188-199.

In preparation:

[1] **L. Wu**, M. Li, J. Wang, and F.-X Wu, "Controllability of complex networks and its applications to biological networks," in preparation, 2018.

Appendix B

Copyright Permissions

Copyright forms of thesis-related publications.

The Institution of **Engineering and Technology**
Michael Faraday House
Six Hills Way, Stevenage
Hertfordshire, SG1 2AY
United Kingdom
T +44 (0)1438 313311
F +44 (0)1438 765526
www.theiet.org

Ref: LWU/Permission/SYB.0077

FAO:

Contact: Lin Wu
Division of Biomedical Engineering, Engineering Building, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK S7N 5A9, Canada
Email: liw557@mail.usask.ca
Phone: +1-306-966-2911

1 June 2018

Permission to reproduce IET content

Dear Sirs,

Minimum steering node set of complex networks and its applications to biomolecular networks ("the Material") to be used in the thesis 'Complex network controllability and applications to biomolecular networks' by Lin Wu, to be published under the University of Saskatchewan ("the Work")

The Institution of Engineering and Technology ("the IET") hereby grants to Lin Wu ("LWU") non-exclusive permission to use the Material in the Work subject to the terms set out below:

Territory:	Worldwide
Languages:	All languages
Term:	Life of the Work
Media:	All print and electronic media how known or hereafter devised
Credit:	Wu, L., Li, M., Wang, J., et al.: 'Minimum steering node set of complex networks and its applications to biomolecular networks', IET Systems Biology, 2015, 10, (3), pp. 116-123
Fee:	None

The permission granted by this letter may not be licensed or assigned without the prior written consent of the IET.

Neither party shall be liable to the other for indirect, incidental, special or consequential damages arising out of or in relation to this agreement (unless such liability cannot be excluded or limited by law).

The parties agree that this letter constitutes the entire agreement between them relating to the use of the Material by LWU and supersedes all previous agreements, understandings and arrangements between them, whether in writing or oral in respect of its subject matter.

No variation of this agreement shall be valid or effective unless it is in writing, refers to this agreement and is duly signed or executed by, or on behalf of, each party.

This letter and any dispute or claim arising out of, or in connection with, it, its subject matter or formation (including non-contractual disputes or claims) shall be governed by, and construed in accordance with, the laws of England and Wales and the parties irrevocably agree that the courts of England and Wales shall have exclusive jurisdiction to settle any

dispute or claim arising out of, or in connection with, this letter, its subject matter or formation (including non-contractual disputes or claims).

Please sign and return the enclosed copy of this letter to confirm your agreement to it.

Yours faithfully

James Sutherland
Permissions Officer

We confirm our agreement to the terms set out above

Signed: ……
Print name: ……Lin Wu……

Date: Jun. 1, 2018

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Jun 06, 2018

This Agreement between University of Saskatchewan -- Lin Wu ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	4363040190805
License date	Jun 06, 2018
Licensed Content Publisher	Oxford University Press
Licensed Content Publication	Bioinformatics
Licensed Content Title	CytoCtrlAnalyser: a Cytoscape app for biomolecular network controllability analysis
Licensed Content Author	Wu, Lin; Li, Min
Licensed Content Date	Nov 23, 2017
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Complex network controllability and applications to biomolecular networks
Publisher of your work	University of Saskatchewan
Expected publication date	Aug 2018
Permissions cost	0.00 CAD
Value added tax	0.00 CAD
Total	**0.00 CAD**
Title	Complex network controllability and applications to biomolecular networks
Instructor name	Fang-Xiang Wu
Institution name	University of Saskatchewan
Expected presentation date	Aug 2018
Order reference number	4351470335961
Requestor Location	University of Saskatchewan 57 Campus Drive University of Saskatchewan Saskatoon, SK S7N 5A9 Canada Attn: University of Saskatchewan
Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	University of Saskatchewan 57 Campus Drive University of Saskatchewan Saskatoon, SK S7N 5A9

Canada
Attn: University of Saskatchewan

Total **0.00 CAD**

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.
5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.
6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com
7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.
8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.
9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.
10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
12. Other Terms and Conditions:
v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.